AJUR

American Journal of *Undergraduate Research*

Volume 22 | Issue 4 | December 2025

www.ajuronline.org

Print Edition ISSN 1536-4585
Online Edition ISSN 2375-8732

AJUR
American Journal of
Undergraduate Research

Volume 22 | Issue 4 | December 2025 | *https://doi.org/10.33697/ajur.2025.155*

American Journal of Undergraduate Research (AJUR) *is a national, independent, peer-reviewed, open-source, quarterly, multidisciplinary student research journal. Each manuscript of AJUR receives a DOI number. AJUR is archived by the US Library of Congress. AJUR was established in 2002, incorporated as a cheritable not-for-profit organization in 2018. AJUR is indexed internationally by EBSCO and Crossref with ISSNs of 1536-4585 (print) and 2375-8732 (web).*

Fraud Detection in Incentive-Based Online Surveys Through Follow-up Demographic Verification

Andy E. Lewis[*a], Kevin B. Gittner*[ab], & Lauren M. Matheny*[b]

[a]*Department of Health Promotion and Physical Education, Kennesaw State University, Kennesaw, GA*
[b]*School of Data Science and Analytics, Kennesaw State University, Kennesaw GA*

https://doi.org/10.33697/ajur.2025.156

*Student: alewi240@students.kennesaw.edu**
Mentors: kgittner@kennesaw.edu, lmathen1@kennesaw.edu

ABSTRACT
Online survey platforms like Qualtrics, SurveyMonkey, and Amazon Mechanical Turk have become essential tools for fast, cost-effective data collection, but growing concerns over data quality and the risk of fraud have accompanied their rapid growth. The purpose of this study was to determine whether follow-up demographic verification surveys can identify fraudulent participation and improve data quality in survey research conducted through online survey platforms.
An orthopedic activity level survey with 35 embedded data quality checks was distributed using Amazon Mechanical Turk. A follow-up demographic verification survey was sent to 28 participants who contacted the survey administrator after completing the original survey. Five datasets were cleaned and merged using R, allowing for the identification of potential patterns of fraud through the direct comparison of demographic information between the original and follow-up surveys.
Participants who did not complete the follow-up survey exhibited signs of fraud and poor data quality, including one email being linked to multiple participant accounts and higher counts of failed data quality checks. Although the small sample size limited the ability to detect statistical significance, descriptive patterns indicate practically meaningful differences. In contrast, demographic discrepancies were minimal among those who completed the follow-up. Open-text box similarity detection was the most effective individual data quality check.
Integrating follow-up verification surveys into study designs provides a practical and scalable approach to detect fraud and maintain data integrity before compensation is distributed, offering researchers a cost-effective method to protect incentive budgets and data quality within online survey research.

KEYWORDS
Online Survey Platforms; Data Quality; Online Research; Fraud Detection; MTurk; Participant Verification; Demographic Consistency; Incentive Structures; Follow-Up Surveys; Crowdsourced Data

INTRODUCTION

Online surveys play a pivotal role in shaping decisions, and the industry's current value and projected growth demonstrate their widespread adoption and expanding role. Already valued at an estimated three billion dollars in 2024, the online survey industry is expected to continue to grow, projected to reach $36 billion by 2030.[1, 2] Online survey platforms such as Momentive (herein referred to as SurveyMonkey), Amazon Mechanical Turk (MTurk), and Qualtrics allow researchers and other professionals across a broad spectrum of fields to collect and harness data in a quicker, cost-effective, and accessible manner than traditional methods.[3–5]

These platforms have extended their reach into public, private, and academic sectors. Regardless of the field or sector, obtaining good-quality data is integral to making valid conclusions that inform decision-making. Despite these advantages, online survey platforms introduce risks of fraud and careless responses, raising concerns among researchers about threats to data quality and research outcomes.[3, 6, 7] Financial incentives can further motivate bad actors to prioritize quantity over quality, leading to fraudulent activities in survey responses that waste valuable resources and distort study results.[8–10]

Survey incentives have been linked to an increase in multiple submissions,[11, 12] and while strategies have been developed to mitigate fraudulent participation, many remain ineffective against those who continue to adapt their techniques to exploit incentive structures. To combat survey fraud, researchers have explored altering incentives, verifying demographic details to identify repeated attempts, and delaying or withholding compensation until fraud detection checks are complete.[9, 13–15] However, the effectiveness of these strategies remains uncertain and requires further evaluation. Without effective deterrents, researchers

risk allocating limited resources to unreliable data. While artificial intelligence (AI) and deep learning offer tools for detecting fraudulent survey responses, both automated and human-generated, their practical use remains limited and not yet widely validated.[16, 17] At the same time, AI lowers the technical barriers for fraudulent actors to exploit surveys. These tools enable automated submissions and allow individuals lacking proficiency in the survey language to use large language models for translation, compromising instruments validated for linguistically proficient respondents and complicating detection of invalid responses.[17, 18]

Given these ongoing challenges, the purpose of this study was to assess and identify fraudulent responses using a follow-up demographic verification survey prior to incentive disbursement. In addition to the embedded data quality checks in the original survey, the follow-up survey functioned as a secondary fraud detection step for a subset of participants who contacted the survey administrator. Instead of relying solely on post-hoc fraud detection methods that can only be used after survey completion and participant compensation, the goal of the follow-up survey was to confirm participant identities through time-delayed verification of reported static demographics and identify discrepancies with their initially reported demographic information. By verifying demographic consistency, this study employs a scalable solution for detecting inconsistencies before incentives are distributed, thereby reducing fraudulent claims, conserving research resources, and improving data quality.

METHODS AND PROCEDURES

Survey design and data collection

An orthopedic activity level assessment survey was distributed in the Summer of 2024 through MTurk. This study was approved by the Kennesaw State University Institutional Review Board (IRB-FY23-373). To examine data quality, 35 quality checks were embedded into the original survey. These checks included logic, demographic consistency, and open-text similarity checks, among others. Participants entered the study with a unique MTurk participant ID. The original survey was divided into two parts, with part one including a set of screening questions and part two containing the primary content for those who qualified. Upon completing the original survey, participants received a survey completion code. This code was required to receive payment through MTurk and was used to verify participation.

Following the completion of the original survey, the data cleaning process was initiated to identify potential fraud. During this time, some respondents contacted the survey administrator about incentive payments, at which point they were sent a follow-up demographic verification survey. These individuals were the only participants to receive the follow-up survey, as the purpose was to confirm their identity before addressing their concerns or distributing compensation. This approach provided a necessary verification step, but it limited the follow-up survey to a self-selected subset, potentially introducing bias. To assess data quality differences, the average number of failed data quality checks in the original survey was compared between those who completed the follow-up survey and those who did not complete it after reaching out to the survey administrator.

Data cleaning and merging process

All data cleaning, handling, and analysis were conducted in R version 4.3.1 using RStudio.[19] Data cleaning and initial quality assessments focused on identifying discrepancies, including missing participant IDs, duplicate entries, and incorrect survey completion codes. These inconsistencies, combined with extensive fraud, made a straightforward merge between datasets impossible. Survey outputs were examined to align participant IDs, survey completion codes, and email confirmation codes used to link the original survey data with the follow-up survey response data. A frequency table was generated to detect duplicate IDs, while other inconsistencies were resolved through manual validation and recoding only when necessary. Email communications were referenced to match records in cases where participant IDs and survey completion codes were missing or incorrect.

Five datasets were imported into RStudio and merged for analysis:

A. Part one of the original survey output (inclusion criteria screening survey)
B. Part two of the original survey output (primary survey)
C. An MTurk file linking each participant to a submission in the original survey, including each Amazon Mechanical Turk participant's unique identifier and their corresponding survey completion code
D. A file containing messages from participants who contacted the survey administrator
E. The follow-up demographic verification survey output

Preparation for analysis involved a series of merges, using both inner and outer joins, to integrate all relevant data into a single dataset (**Figure 1**). First, part one (A) and part two (B) of the original survey output were merged, then linked to MTurk records (C). Duplicated responses were removed, producing a cleaned dataset of unique responses from the original survey. Next, messages from participants who contacted the survey administrator (D) were merged with the follow-up survey responses (E), matching individuals who had reached out to the survey administrator with their follow-up data. Finally, the follow-up survey dataset (D & E) was merged with the cleaned original survey and MTurk records (A, B, & C), creating a single dataset (n = 28)

with unique identifiers. The final dataset was used to validate participant demographics between the original and follow-up survey responses, confirming whether respondents reported the same residential ZIP code, birth year, and email address.

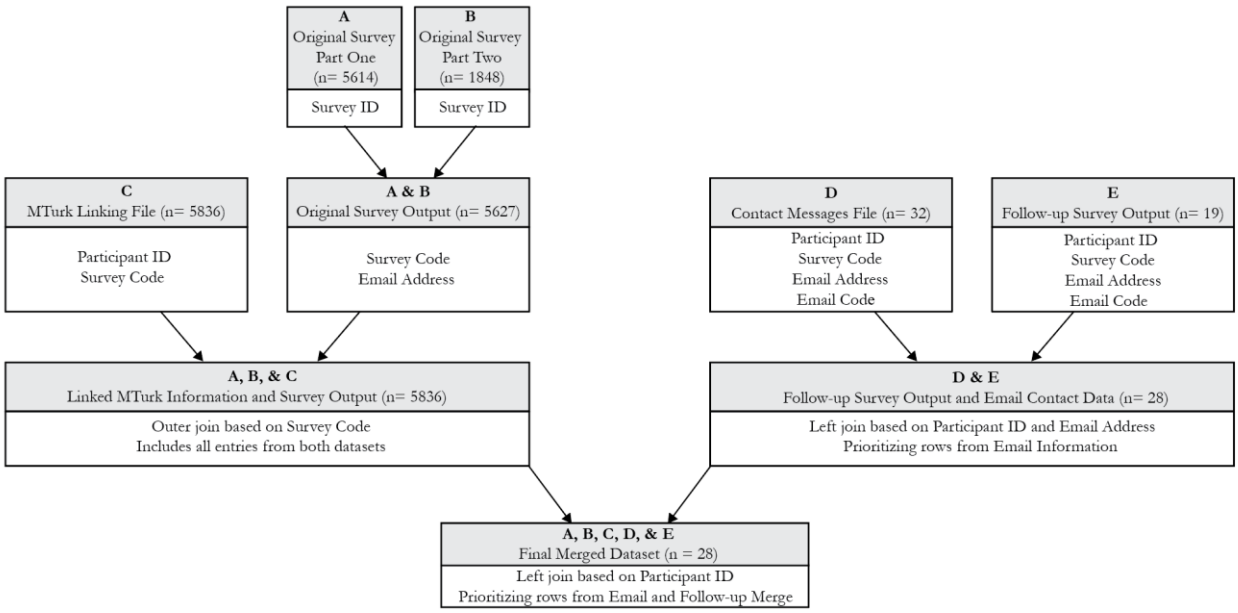


Figure 1. Merging steps. Five datasets combined into a single file for analysis.

Statistical analysis

Descriptive statistics were calculated to summarize demographic consistency checks and original survey quality check failures. Deidentified patterns were visualized to summarize trends observed after the merging process. The mean number of failed embedded data quality checks in the original survey was used as an indicator of data quality and was compared between participants who completed the follow-up survey and those who did not. Histograms of residuals, QQ-plots, and a Shapiro-Wilk test were used to assess the normality of the distribution of counts of failed data quality checks from the original survey for participants who completed the follow-up survey and those who did not. Mean and median counts of failed data quality checks were also compared to assess potential differences between the groups. A Mann-Whitney U test was conducted to determine whether there was a statistically significant difference in the distribution of the number of failed data quality checks between participants who completed the follow-up and those who did not. Finally, summary statistics were used to evaluate the effectiveness of the original survey's 35 individual data quality checks, which were designed a priori to identify the most effective checks.

RESULTS

Demographic comparison

After removing duplicates, a total of 28 emails were sent to the survey administrator. The demographic verification survey was sent to all 28 participants, of whom 53.6% completed the follow-up survey. Among those who completed the follow-up, only minimal demographic discrepancies were observed. Specifically, there was one ZIP code, two birth year, and two email inconsistencies between the original survey and the follow-up survey. Although these inconsistencies raise slight concerns about data reliability from this group, the issues were minor compared to the discrepancies observed among participants who did not complete the follow-up survey. Demographic consistency validation revealed that participants who completed the follow-up survey (n = 15) showed consistency across ZIP codes, birth years, and email addresses, whereas participants who did not complete the follow-up survey (n = 13) demonstrated substantial email inconsistencies (**Table 1**).

Consistency Check	Completed follow-up (n = 15)	Did not Complete follow-up (n = 13)
ZIP Code	14/15 (93.3%)	—
Birth Year	13/15 (86.7%)	—
Email	13/15 (86.7%)	3/13 (23.1%)

Table 1. Demographic consistency across participants who completed and did not complete the follow-up survey. For participants who completed the follow-up survey, demographic consistency was assessed by comparing the ZIP Code, birth year, and email reported in the follow-up survey with those in the original survey. For participants who did not complete the follow-up, email consistency was assessed by comparing the email used to contact administrator with email used in the original survey.

Among participants who did not complete the follow-up survey (n = 13), 76.9% demonstrated email inconsistencies (**Table 1**), including six respondents who were all linked to a single email address across six separate MTurk worker accounts (**Table 2**). This pattern suggests that one individual may have taken the survey multiple times under different identities, with the time-delayed verification introducing recall challenges that reduced the consistency of fraudulent responses.

Participant (n = 13)	Contact Email	Original Survey Email	Email Consistency?
Participant ID 1	Contact-A@example.com	**Survey-A@example.com**	×
Participant ID 2	Contact-B@example.com	**Survey-A@example.com**	×
Participant ID 3	Contact-C@example.com	**Survey-A@example.com**	×
Participant ID 4	Contact-D@example.com	**Survey-A@example.com**	×
Participant ID 5	Contact-E@example.com	**Survey-A@example.com**	×
Participant ID 6	Contact-F@example.com	**Survey-A@example.com**	×
Participant ID 7	Contact-G@example.com	Survey-B@example.com	×
Participant ID 8	Contact-H@example.com	Survey-C@example.com	×
Participant ID 9	Contact-I@example.com	Survey-D@example.com	×
Participant ID 10	Contact-J@example.com	Survey-E@example.com	×
Participant ID 11	Contact-K@example.com	Contact-K@example.com	✓
Participant ID 12	Contact-L@example.com	Contact-L@example.com	✓
Participant ID 13	Contact-M@example.com	Contact-M@example.com	✓

Table 2. Email inconsistencies among those who did not complete the follow-up survey (n = 13). ✓ indicates that the contact email used to reach the survey administrator matched the email provided in the original survey. × indicates a mismatch between the contact email and the original survey email. The bolded example email (e.g., Survey-A@example.com) was used across multiple unique participant accounts.

Descriptive analysis

To assess whether there were observable differences in data quality, the counts of failed data quality checks from the original survey were compared between participants who completed the follow-up survey and those who did not. Participants who did not complete the follow-up survey had higher mean and median counts of failed data quality checks in the original survey (mean = 5.9, SE = .84, median = 6), compared to those who did complete the follow-up survey (mean = 4.0, SE = 1.03, median = 2.5). Normality assessments indicated a violation of normality in the group that completed the follow-up survey ($W = 0.797$, $p = 0.005$). Given the violation of normality and the small sample size, non-parametric measures were used to compare the counts of failed data quality checks in the original survey between participants who completed the follow-up survey and those who did not. The Mann-Whitney U test found no significant difference in failed quality checks between groups ($U = 125$, $p = 0.101$). Given the small sample size of this study, the analysis was underpowered, limiting the ability to detect group differences. However, higher counts of failed data quality checks among those who did not complete the follow-up suggest a trend of poorer data quality, while clustering of high-quality check failures among some participants that completed the follow-up survey demonstrates that layered detection approaches remain necessary (**Figure 2**).

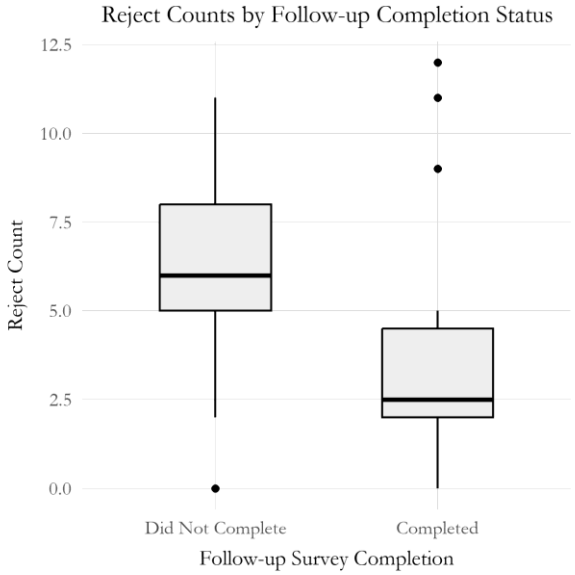


Figure 2. Side-by-side boxplot of failed data quality checks in the original survey. Comparisons are made between follow-up survey non-completion and completion groups.

Effectiveness of specific data quality checks

A deeper analysis of the 28 follow-up participants revealed that certain data quality checks were more effective at identifying poor data quality in the original survey. Among these, half of the responses failed data quality checks due to email-related issues. Specifically, seven participants had different email confirmation codes reported than the one sent to them, and seven had duplicated email addresses used across multiple submissions. An additional six cases involved the participants' provided age not matching the provided year born with a two-year margin of error. Another six responses failed checks because the first seven words of one open-ended question regarding ankle function were identical to those in another response. Additionally, 14 different cases failed the same seven word similarity check for a separate open-ended question about honesty. Both open-ended checks failed participants whose answers matched text from other responses in the original survey. These counts of failed data quality checks for similarity detection suggest that some participants may have copied text from other responses or reused their answers across multiple entries.

Eight respondents provided a residential ZIP code that was inconsistent with their reported state of residence. In logical comparison data quality control checks, five failed checks related to task difficulty (e.g., reporting that getting out of a tub was more difficult than running a marathon). Twelve respondents failed checks for incorrectly identifying how they received the original survey. Although all participants were recruited through MTurk, they selected options such as Facebook, Instagram, or word of mouth instead of the correct answer, "Other." Additionally, four respondents failed a response strategy check for selecting "careless" or "inattentive" when asked to indicate their strategy for responding to the survey when they should have selected "Other." **Table 3** summarizes the number of respondents that failed each data quality check out of the total sample of 28.

Data Quality Check	Number of Failed Data Quality Checks	Percentage of Participants with Failed Checks (n = 28)
Open-ended honesty similarity checks	14	50%
Survey recruitment method check	12	42.8%
State vs ZIP Code check	8	28.6%
Email code check	7	25%
Duplicate email check	7	25%
Age check	6	21.4%
Open-ended ankle similarity check	6	21.4%
Tub and marathon check	5	17.9%
Response strategy check	4	14.3%

Table 3. Counts and percentages of respondents failing data quality checks (n = 28).

DISCUSSION

Among those who contacted the survey administrator, similarity detection in open-ended responses proved to be the most effective in identifying poor data quality. Participants who did not complete the follow-up survey showed higher reject counts compared to those who did. While this finding was not statistically significant, results indicate a practically meaningful trend that may extend to other online survey platforms, suggesting that participants who avoid verification steps may disproportionately contribute to fraudulent or low-quality data. Replicating findings in a larger sample size would provide the statistical power needed to detect group differences and better evaluate whether the descriptive trend reflects both a practical and statistically significant effect.

Although these findings were observed in MTurk data, concerns about inattentive or fraudulent participation are not exclusive to this platform, and there is no clear consensus on which platform produces the best data. While some studies suggest that Prolific produces higher-quality data than MTurk or Qualtrics,[6, 8] no platform is immune to fraud or poor-quality responses, making verification methods, such as follow-up demographic checks, a solution that can be applied regardless of the platform. One way to scale this approach is to make completion of a delayed time follow-up demographic verification survey a condition for receiving incentives in any study conducted through online survey platforms.

Age and ZIP code consistency could not be fully validated for participants who did not complete the follow-up survey. Analysis was restricted to comparing contact emails with emails from the original survey, limiting the ability to assess fraudulent behavior entirely. This limitation also serves as a key finding, as the high rate of non-response may itself indicate intentional avoidance of verification when dishonest participants are presented with verification before receiving incentives. The follow-up survey was sent after the original survey in response to participant contact, allowing those who did not complete it to opt out of the verification process. This self-selected follow-up group should be interpreted with caution, given the potential for selection bias. Finally, because this study was conducted exclusively on MTurk with a small, nonrandom sample, findings may not be generalizable to other online survey platforms or broader populations.

The follow-up survey enabled the identification of demographic inconsistencies, allowing for the detection of potential fraudulent and dishonest participation. Notably, many of those who did not complete the follow-up survey exhibited suspicious behavior, with nearly half having multiple MTurk participant accounts linked to a single email within the original survey. This supports and expands on other research that examined the use of follow-up surveys as a secondary point of contact, such as demographic verification, as a method to validate participant identity, detect fraud, and reduce fraudulent participation.[9, 14, 15] The results of this study demonstrate how such methods can identify similar deliberate attempts at providing dishonest information, as reported demographic information such as ZIP code, year of birth, or email address should not change in delayed short-term follow-up requests.

Researchers can implement similar follow-up verification procedures within their study planning to serve as an additional fraud detection mechanism after the completion of primary data collection, allowing for a delayed evaluation of respondent legitimacy. This approach is a valuable protective measure for researchers in managing incentive distribution efficiently, as verification prior to payout can reduce unnecessary costs associated with fraudulent claims. This study demonstrates how even non-expert researchers can adopt manual follow-up procedures to detect fraud.

Automation could streamline verification by flagging inconsistent demographic responses, eliminating the need for manual data cleaning. Future research could explore the use of machine learning techniques to automate the detection of repeated responses in open-ended questions. This would improve scalability and reduce manual labor in larger datasets. By incorporating follow-up verification surveys into the research design, investigators can improve data quality while reducing fraudulent claims that exploit survey incentives. As fraud detection methods continue to evolve, integrating multi-layered verification strategies will be essential for maintaining data integrity in online research.

CONCLUSIONS

High-quality data collection through online survey platforms continues to be a challenge, as incentives can increase vulnerability to fraudulent exploitation. When researchers lack a defined sampling frame and are unable to verify participant identity at baseline, a common limitation in online data collection, linking incentives to a follow-up verification step provides a practical and scalable strategy to reduce fraud and preserve data quality. As fraud tactics continue to evolve alongside advancements in artificial intelligence, verification measures should adapt in parallel. This study confirmed that a single individual attempted to manipulate the incentive system by contacting the research team under multiple participant IDs tied to the same survey email address in the original survey. This finding highlights the very real risk of deliberate fraud in online data collection platforms. Future large-scale studies are needed to assess this behavior further and refine follow-up verification as a countermeasure.

ACKNOWLEDGEMENTS

The student author thanks the members of the Kennesaw State University's Data Quality and Survey Methods Lab for their assistance throughout the project. Additional thanks go to Kennesaw State University for providing the resources and environment that made this work possible.

REFERENCES

1. The Business Research Company, Online Survey Software Global Market Report 2025, *https://www.thebusinessresearchcompany.com/report/online-survey-software-global-market-report* (accessed Jan 2025)
2. Knowledge Sourcing Intelligence LLP, Online Survey Software Market Size: Industry Report, 2023–2028, *https://www.knowledge-sourcing.com/report/global-online-survey-software-market* (accessed Jan 2025)
3. Newman, A., Bavik, Y. L., Mount, M., and Shao, B. (2021) Data collection via online platforms: challenges and recommendations for future research, *Appl Psychol Health Well Being* 70(3), 1381–1387. *https://doi.org/10.1111/apps.12302*
4. Callegaro, M., Manfreda, K., and Vehovar, V. (2015) Survey Research and Web surveys, in *Web Survey Methodology* (Metzler, K., Ed.) 1st ed., 1–34, SAGE Publications Ltd., London. *https://doi.org/10.4135/9781529799651*
5. Agans, J. P., Schade, S. A., Hanna, S. R., Chiang, S. C., Shirzad, K., and Bai, S. (2024) The inaccuracy of data from online surveys: a cautionary analysis. *Qual Quant* 58(3), 2065–2068. *https://doi.org/10.1007/s11135-023-01733-5*
6. Douglas, B. D., Ewell, P. J., and Brauer, M. (2023) Data quality in online human-subjects research: comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA, *PLOS ONE* 18(3) e0279720, 1–5. *https://doi.org/10.1371/journal.pone.0279720*
7. Dupuis, M., Renaud, K., and Searle, R. (2022) Crowdsourcing quality concerns: an examination of Amazon's Mechanical Turk, *Proceedings of the 23rd Annual Conference on Information Technology Education*, 127–129. *https://doi.org/10.1145/3537674.3555783*
8. Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2022) Data quality of platforms and panels for online behavioral research, *Behav Res* 54, 1643–1653. *https://doi.org/10.3758/s13428-021-01694-3*

9. Johnson, M. S., Adams, V. M., and Byrne, J. (2023) Addressing fraudulent responses in online surveys: insights from a web-based participatory mapping study, *People Nat (Hoboken)* 6(1), 148–161. *https://doi.org/10.1002/pan3.10557*
10. Kumarasamy, V., Goodfellow, N., Ferron, E. M., and Wright, A. L. (2024) Evaluating the problem of fraudulent participants in health care research: multimethod pilot study, *JMIR Form Res* 8 e51530, 2–10. *https://doi.org/10.2196/51530*
11. Chesney, T., and Penny, K. (2013) The impact of repeated lying on survey results, *SAGE Open* 3(1), 1–2. *https://doi.org/10.1177/2158244012472345*
12. Langer, A., Leopold, L., Meuleman, B., Kuppens, L., and Menon, A. (2025) Mitigating fraud in incentivized online surveys: lessons from Facebook recruitment in Nigeria, *Surv Method: Insights Field* 3(2). 4–7. *https://doi.org/10.13094/SMIF-2025-00006*
13. Johns Hopkins University, Guidance on Fraud Prevention Regarding Use of Survey Instruments, Version 1, *https://publichealth.jhu.edu/sites/default/files/2025-01/Fraud-Prevention-Regarding-Use-of-Survey-Instruments.pdf* (accessed Jan 2025)
14. Teitcher, J. E. F., Bockting, W. O., Bauermeister, J. A., Hoefer, C. J., Miner, M. H., & Klitzman, R. L. (2015) Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs, *J Law Med Ethics* 43(1), 119–128. *https://doi.org/10.1111/jlme.12200*
15. Guy, A. A., Murphy, M. J., Zelaya, D. G., Kahler, C. W., and Sun, S. (2024) Data integrity in an online world: demonstration of multimodal bot screening tools and considerations for preserving data integrity in two online social and behavioral research studies with marginalized populations, *Psychol Methods,* 14–16. *https://doi.org/10.1037/met0000696*
16. NORC. Promise and Pitfalls of AI-Augmented Survey Research, *https://www.norc.org/research/library/promise-pitfalls-ai-augmented-survey-research.html* (accessed Sep 2025)
17. Dynata. The Double-Edged Sword of Generative AI: Mitigating the Risks of Respondent Fraud, *https://www.dynata.com/why-dynata/resources/blog/the-double-edged-sword-of-generative-ai-mitigating-the-risks-of-respondent-fraud/* (accessed Sep 2025)
18. Lebrun, B., Temtsin, S., Vonasch, A., and Bartneck, C. (2024) Detecting the corruption of online questionnaires by artificial intelligence, *Front Robot AI* 10, 1–10. *https://doi.org/10.3389/frobt.2023.1277635*
19. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, *https://www.R-project.org/* (accessed Feb 2025)

ABOUT THE STUDENT AUTHOR

Andy Lewis graduated from Kennesaw State University in Spring 2025 with a Bachelor of Science in Integrated Health Science. He will be continuing his education at Kennesaw State as a graduate student in the Master's in Data Science and Analytics program. This project served as an important stepping stone in that journey, combining his interests in research, public health, and data analysis.

PRESS SUMMARY

Online surveys are a popular tool for researchers but can be vulnerable to fraud. This study investigated whether a short follow-up survey could help confirm participant identity and improve data quality. Researchers initially conducted an orthopedic survey on Amazon Mechanical Turk to evaluate data quality on the platform. Participants who reached out to the survey administrator about compensation were later sent a follow-up survey to verify demographic details. The analysis revealed signs of possible dishonesty, including one email linked to multiple participant accounts and other inconsistencies. The findings suggest that follow-up verification surveys can help protect data quality and research budgets.

Siderophore Production by a Rhizosphere-Associated *Streptomyces* from *Cyperus virens*

Phoebe Dennison[a], Robert Samples[b], Riccardo Racicot[b], & Lesley-Ann Giddings[a,b]*

[a]Department of Chemistry, Smith College, Northampton, MA
[b]Biochemistry Program at Smith College, Northampton, MA

https://doi.org/10.33697/ajur.2025.158

*Student: pdennison@smith.edu**
Mentor: lgiddings@smith.edu

ABSTRACT
Rhizosphere microbiomes produce iron chelators or siderophores to capture ferric iron, an essential nutrient for growth that is not always bioavailable. Plants with siderophore activity aid in phytoremediation by removing heavy metals and other pollutants. The *Cyperus* genus of plants has a high affinity for iron uptake in the phytoremediation of wastewater. Herein, we isolated a siderophore-producing microbe from the rhizosphere of *Cyperus virens*, an understudied member of this genus, using the Chrome Azurol S (CAS) assay. While this isolate has siderophore activity, it did not exhibit antimicrobial activity when plated on *Aspergillus niger, A. flavus*, or *Bacillus subtilis*. The isolate was identified as *Streptomyces* sp. PD-S100-1 through genomic sequencing and de novo assembly with a draft genome size of 8.2 Mbp and 73% GC content. antiSMASH analysis of the genome identified several siderophore biosynthetic gene clusters, including those involved in the production of mirubactin A and bacillibactin. Liquid chromatography-mass spectrometry (LC-MS) detected several siderophores, including mirubactin A–D, enterobactin, and other catecholates from these siderophore families. The positive CAS assay, siderophore gene cluster identification, and LC-MS/MS analyses show that the rhizosphere of *C. virens* contains siderophore-producing bacteria. Most detected metabolites, including enterobactin, increased in the presence of cerium, a lanthanide involved in the expression of secondary metabolites, whereas mirubactin production was reduced. The presence of rhizosphere siderophore-producing bacteria suggests that *C. virens* may have potential applications in environmental phytoremediation, targeting pollutants from wastewater, mining, agriculture, and energy production.

KEYWORDS
Siderophore; Mirubactin A; Enterobactin; Bacillibactin; Rhizosphere; *Cyperus virens*; *Streptomyces*; CAS Assay

INTRODUCTION

Iron is an essential nutrient for plants and microorganisms, but its bioavailability is highly oxidation state-dependent. Ferrous iron (Fe^{2+}) is more bioavailable, while ferric iron (Fe^{3+}) has greater biochemical importance, but is only water-soluble at concentrations up to 10^{-9} M.[1] Most bacteria require 10^{-6} M of intracellular Fe^{3+} for viability and produce siderophores, low molecular weight iron-chelating molecules that mediate ferric iron uptake.[1] There are four main types of siderophores: hydroxamates, catecholates, carboxylates, as well as numerous mixed-type siderophores with various metal binding moieties.[1] The most common class of siderophores in nature is the hydroxamates (e.g., ferrioxamine in **Figure 1)**, which are produced by both bacteria and fungi.[2]

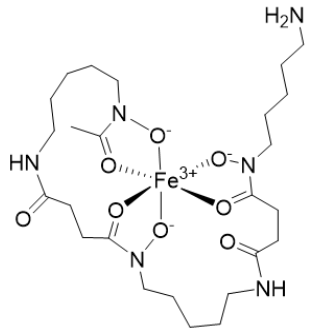


Figure 1. Structure of the hydroxamate siderophore ferrioxamine in complex with Fe^{3+}.

The metal-chelating capabilities of siderophores have applications in medicine, agriculture, and environmental protection. Siderophore-producing plants and microbes can be used to remediate natural or anthropogenic heavy metal contamination by

binding metals, reducing their bioavailability in soil.[1] These bacteria can be found in the plant rhizosphere to help regulate plant growth. Actinobacteria are ubiquitous taxa in the rhizosphere and are one of the most important sources of secondary metabolites, including siderophores.[3] For example, Wang *et al.* found that 44% of actinobacterial rhizosphere isolates produce siderophores.[3]

The application of siderophores in environmental protection through phytoremediation is of utmost importance today, with the continuing development and usage of natural land by humans. The misuse of land and waterways can lead to environmental damage and health issues. To ameliorate the negative effects of heavy metal and oil pollution, the use of plants in phytoremediation is a cost-effective and less complicated strategy than traditional methods, such as excavation, filtration, and adsorption.[4,5] Members of the *Cyperus* genus have been studied for their uptake of heavy metals in the environment. For example, *Cyperus esculentus* can grow in wetlands of wastewater and accumulate heavy metals in its roots with the highest affinity for iron uptake.[6] Understanding more about the siderophore activity of plants to undo or prevent environmental damage is pertinent.

C. virens is a grass-like sedge plant (**Figure 2**) that has not been explored for phytoremediation but holds unconsidered potential as other genus members have successfully been applied in phytoremediation.[7] It is typically found in the southeastern states of the US in moist places like marshes and globally in tropical and subtropical areas of the Americas.[8–10] Within its larger family of *Cyperaceae*, specifically of *Costularia*, 52% of isolated bacterial strains were reported to produce siderophores.[9] Additionally, *Cyperaceae* are commonly found in nutrient-deficient environments,[1] suggesting that the presence of siderophore-producing bacteria is essential for the growth of plants in this family, including *C. virens*. No siderophore activity has been reported from *C. virens* rhizosphere bacteria, making this study an important addition to understanding more about this specific plant and the *Cyperus* genus.

Figure 2. The *C. virens* plant from the Lyman Plant House at Smith College in Northampton, MA.

We hypothesized that the *C. virens* rhizosphere contains actinobacteria that produce siderophores that may be related to the host plant's phytoremediation capabilities. These rhizosphere bacteria may also produce higher yields, a greater diversity, or chemically unique siderophores. To determine siderophore activity in the rhizosphere of *C. virens,* microbial isolation was performed using glucose-yeast-malt (GYM) agar and marine agar to enrich for actinobacteria, which were then analyzed for the presence of siderophores using a Chrome Azurol S (CAS) assay. A draft genome was obtained for an isolate to identify specific gene clusters related to siderophores and other bioactive properties. Liquid chromatography-mass spectrometry (LC-MS) analyses of the isolate grown in the presence and absence of cerium, a strong inducer of secondary metabolites, were performed to detect changes in siderophore production and gain insight into the regulation of their biosynthesis.[11]

METHODS AND PROCEDURES

Plant Sampling and Processing

The roots and rhizosphere of *C. virens* were collected from the Lyman Plant House and Conservatory at Smith College in Northampton, Massachusetts. Root pieces (~2 g) were separated from the soil and baked at 50 °C for 30 minutes to select for heat-resistant, Gram-positive actinobacteria. Baked roots were pestled with autoclaved water to make a root solution (0.25 g root/mL). Soil (<1 g) was baked at 50 °C for 30 minutes and then combined with autoclaved water (0.1 g soil/mL) for one hour of rehydration on a slow stirring plate.

Isolation and Cultivation

Root and soil solutions were serially diluted (10^0 up to 10^{-2}) with autoclaved water. Solutions were mixed, centrifuged at 6000 x *g* for three minutes, and spread (100 µL) on solid glucose, yeast extract, and malt extract (GYM) media containing glucose (4 g/L), agar (10 g/L; BD Biosciences, Franklin Park, NY), yeast extract (4 g/L), and malt extract (2 g/L) with nalidixic acid (20 mg/L, NDA) and cycloheximide (80 mg/L, CHX). Plates were incubated for one week at 37 °C. Marine agar plates (BD Biosciences, Franklin Park, NY) were also prepared using the aforementioned dilution series.

Chrome Azurol S, Antibacterial, and Antifungal Assays

CAS dye was prepared according to the procedure by Louden, et. al.[12] The CAS mixture was poured over the isolate on GYM-NDA-CHX and marine agar plates and incubated at room temperature. For antibacterial and antifungal assays, GYM plates were also spread with either of the following: *Bacillus subtilis, Aspergillus flavus,* and *A. niger*. Cell scrapings of isolates were dropped onto these plates, incubated at 37 °C, and analyzed for zones of inhibition.

DNA Extraction, Library Preparation, and Illumina Sequencing

The extraction of genomic DNA (~20 mg) from cells on GYM agar was performed using the ZymoBIOMICS DNA microprep kit per the manufacturer's instructions. Using the 2 x 151 paired base end NEBNext UltraExpress FS Library Prep Kit (Ipswich, MA), a library of the extracted DNA was prepared from the isolate per the manufacturer's instructions. The quality of the library was assessed using the Fragment Analyzer 5200. The concentration of the library was confirmed using the Qubit 4. The DNA library was sequenced at the Center for Molecular Biology at Smith College using an Illumina MiSeq (San Diego) and MiSeq reagent Kit V3 (600-cycle).

De Novo Genome Assembly and Bioinformatic Analysis

The sequenced genome was uploaded to Galaxy (version 24.1, accessed November 2024), and the reads were combined using the concatenate multiple datasets (Galaxy Version 1.4.3, accessed November 2024) tool.[13] Low-quality trims were eliminated using the Trimmomatic (Galaxy Version 0.39+galaxy2, accessed November 2024) tool.[14] The quality of reads was analyzed using the FastQC (Galaxy Version 0.74+galaxy1, accessed November 1, 2024) tool.[15] The data was assembled using Unicycler (Galaxy version 0.5.1+galaxy0, accessed November 2024) and the resulting assembly was submitted to GenBank.[16] Genome assembly quality was assessed by Quast Galaxy Version 5.2.0+galaxy1 (accessed November 3, 2024) tool, which also determined the number of contigs.[17–20] The taxonomic identification of the sample was completed using NCBI BLAST + blastn (Galaxy version 2.14.1+galaxy2, accessed November 2024) tool, where the entire genome assembly was compared against the NCBI Whole-Genome-Shotgun contigs database.[21–23] Furthermore, Sanger sequencing of the partial bacterial 16S rRNA gene using 27F and 1492 primers further confirmed the taxonomic identification.[23] Due to the challenging taxonomic identification of *Streptomyces* via 16S ribosomal RNA gene sequencing, a set of established target genes (23S rRNA, groEL, 16S rRNA, rpoB, recA, gyrB, ssr1A, secA, cox1B, and trpB) along with the largest contig were blasted against the NCBI core NT database for taxonomic identification.[24] The RAST annotation tool was used to visualize related functional genes coded as subsystems within the genome using the RASTtk feature.[25] The final assembled genome was annotated using Prokka (Galaxy Version 1.14.6+galaxy1, accessed November 2024).[26,27] Identification of secondary metabolites was completed using the antiSMASH (Galaxy Version 6.1.1+galaxy1, accessed November 2024) tool with default parameters.[28]

Liquid Cultivation for LC-MS/MS

Liquid GYM media (10 mL) was inoculated with cell scrapings. A cerium edetate stock solution (50 mM) was prepared with equal volumes of aqueous stocks of $CeCl_3 \cdot 7\ H_2O$ (100 mM; Sigma-Aldrich, St. Louis, MO, USA) and disodium ethylenediaminetetraacetic (100 mM; Sigma-Aldrich, St. Louis, MO, USA). Using a 96-well plate, three culture replicates were used for the following condition and incubated for seven days at 30 °C stationary: GYM media (control; 200 µL); GYM (180 µL) with inoculant (20 µL); and GYM (160 µL) with inoculant (20 µL) and NaCe-EDTA (2.5 mM, 20 µL).

Liquid Chromatography/Mass Spectrometry (LC/MS)

LC/MS grade methanol (200 µL) was added to cultures grown in the 96-well plate, sonicated for one minute, centrifuged for five minutes at 6000 x *g*, and analyzed by LC/MS. A Thermo Scientific (Waltham, MA) Q-exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer interfaced with a Vanquish Horizon Ultra High Performance Liquid Chromatography (UHPLC) system and VH-D10-A UV detector using a Waters (Milford, MA) HSS T3 C18 column (1.8 µm, 2.1 × 150 mm) was used for analysis. Using a flow rate of 0.5 mL min-1, samples (2 µL) were injected onto the column with a maintained chamber temperature of 40 °C. UHPLC separation was achieved using the following gradient method: 2% acetonitrile: 98% water with 0.1% formic acid for 1 min, 2–40% over 4 min, 40–98% over 3 min, 98–2% over 0.2 min, and 2% for 2 min. The UV-vis data were collected with 200–300, 300–400, 400–500, and 500–600 nm channels. The following electrospray ionization (ESI) settings were used in probe position D: 40 sheath gas flow, 8 auxiliary gas flow, 1 sweep gas flow, 3.5 kV spray voltage, 380 °C capillary

temperature, 50 radiofrequency (RF) funnel level, and 350 °C auxiliary gas temperature. A Top 5 DDA method was used with MS1 (60,000 resolution) and MS2 (15,000 resolution) scans between 150-2,000 m/z.

LC-MS/MS Data Analysis

For untargeted metabolomic analyses, raw data was processed in MS-DIAL v5.5.241113 with default settings,[29] profile data for MS1 and MS2, [M+H] and [M+Na] adducts selected, a minimum peak picking threshold of 3e5 and peak lists analyzed with MPACT v1.00 r24.11.17.[30] Data was filtered and visualized with default settings in MPACT. Targeted analysis was conducted in Thermo Freestyle 1.8 SP2 and MS/MS spectra matches were confirmed using NIST 23 and/or GNPS reference spectral libraries.[31]

Data Sharing and Nucleotide Accession Numbers

Genome assembly and metadata were submitted to the National Center for Biotechnology Information under BioProject identifier PRJNA1212226.

RESULTS

Colony Growth and Morphology on GYM-NDA-CHX and Marine Media

Bacterial colonies were selected based on their resemblance to actinobacteria on GYM-NDA-CHX media (**Figure 3**). PD-S100-1 on GYM-NDA-CHX media formed small grayish/white colonies with no diffusible pigment. Colonies were convex in elevation with a dry, fuzzy texture—an indication of sporulating ability. PD-S100-1 also grew on high-salinity marine media but changed from beige small colonies to punctiform colonies.

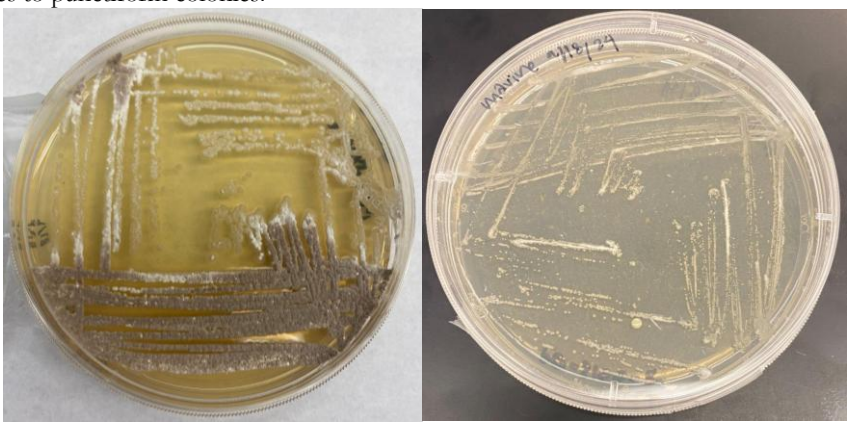


Figure 3. Isolation of PD-S100-1 on GYM-NDA-CHX media (left), showing a higher density of pigmented colonies with increasingly dry texture. PD-S100-1 on high-salinity marine media (right) grows as smaller colonies.

CAS Assay

CAS assays detected the production of siderophores by bacterial colonies. Isolate PD-S100-1 exhibited a density-dependent orange color change on the CAS assay plate, which had the highest colony density (**Figure 4**). The bright orange indicated the isolate was a strong siderophore producer.

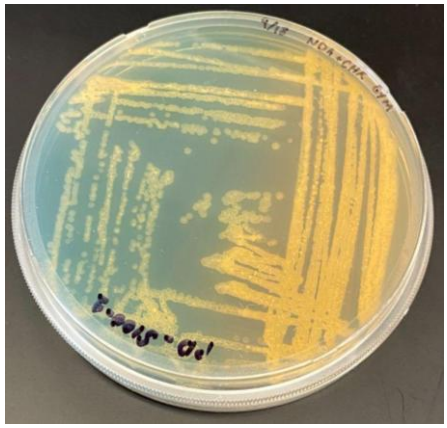


Figure 4. Image of CAS plate assay with isolate PD-S100-1 showing density-dependent siderophore production.

Antifungal and Antibacterial Assays

The antibacterial and antifungal activities were evaluated using *A. flavus*, *A. niger*, and *B. subtilis* on GYM plates. PD-S100-1 showed no antibacterial activity against *B. subtilis*. PD-S100-1 grew on plates inoculated with both of the following fungi: *A. niger* and *A. flavus*.

Genomic Sequencing, Taxonomic Identification, and Secondary Metabolites

A draft genome of PD-S100-1 was obtained and found to be 8.2 Mbp with a 73% GC content. The total number of reads was 1,674,272, and 1,644,916 reads passed quality control. The genome assembly is made up of 261 contigs per the Quast tool and the N50 was 51,839 bp. The isolate was identified as *Streptomyces* sp. When the whole genome and phylogenetic gene markers were compared to NCBI genomic databases. All phylogenetic gene markers were >91% identical to other strains of *Streptomyces*, and the largest contig (30,632 base pairs) was 99.4% identical to *Streptomyces* sp. E14 cont1.653 (See **Supporting Information Tables 1–2**). RAST annotation of the biological functions of this genome (**Figure 5**) revealed that most genes were involved in cellular respiration, phosphorus metabolism, and stress responses. There were also several genes involved in iron acquisition and secondary metabolism.

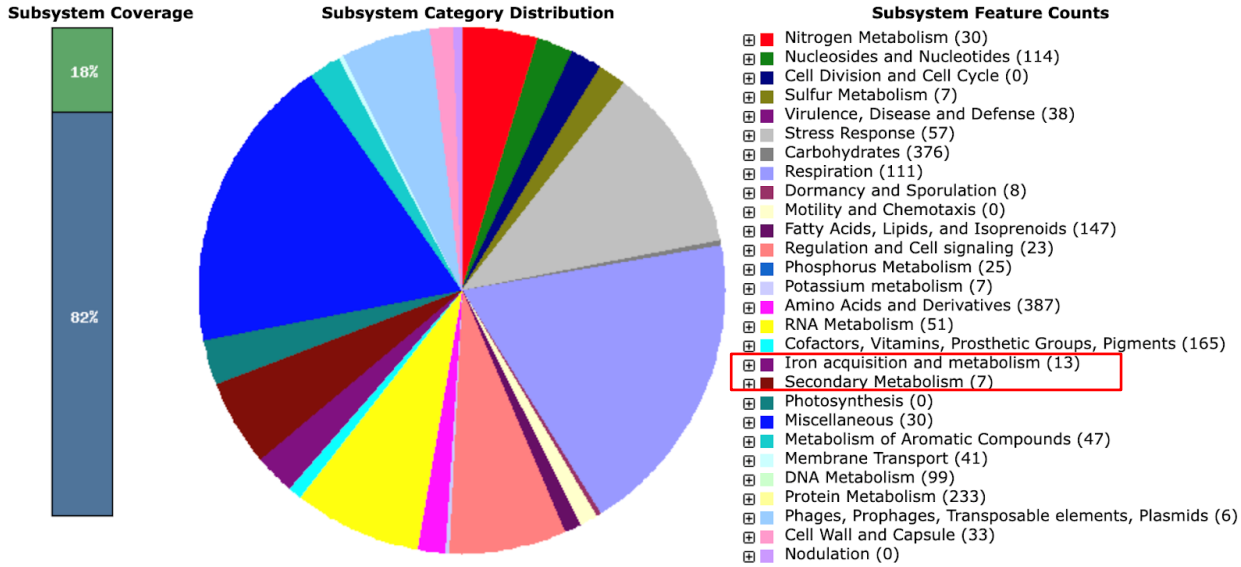


Figure 5. RAST annotation of *Streptomyces* sp. PD-S100-1 genome, indicating related functional roles of genes.[23] The percentage of genes that could be grouped within a subsystem (18%) is shown in green and the percentage of genes that could not be classified (82%) is shown in blue. Feature counts involved in iron acquisition and secondary metabolism are within the red box.

Region	Type	Most Similar Gene Cluster	Similarity
11.1	NRPS	Mirubactin	50%
12.1	Siderophore		
16.1	Terpene	Albaflavenone	100%
21.1	Lanthipeptide-class-iii, RiPP-like	Informatipeptin	100%
51.1	Siderophore		
54.1	NRPS, RiPP-like	Bacillibactin	100%
57.1	Terpene	Hopene	92%
140.1	Terpene	Geosmin	100%
148.1	Ectoine	Ectoine	100%

Table 1. Select identified biosynthetic gene clusters from genomic sequencing of isolate PD-S100-1. 'Region' indicates the string of predicted Pfam domains fed into a hidden Markov model within the antiSMASH pipeline based on the location of the gene cluster within the genome. 'Type' and 'Most Similar Gene Cluster' indicate the class of secondary metabolite and the highest ranked gene cluster, respectively. The ranking system is a composite score based on the number of query and core query genes with a significant hit, gene pairs with conserved synteny as well as those involving a core gene as well as the presence or absence of a core gene. The similarity percentage is the percentage of genes within the known gene cluster with some match to at least one gene within the query region. Only the most similar gene clusters with at least 50% similarity to known gene clusters were included.

antiSMASH identified and annotated twenty-eight secondary metabolite gene clusters in the PD-S100-1 genome. These gene clusters are involved in the production of diverse classes of natural products, including terpenes, siderophores, nonribosomal peptides (NRPs), polyketides, lanthipeptides, and ribosomally synthesized and posttranslationally modified peptide products (RiPPs) with varying degrees of similarity to the sequenced genome of PD-S100-1 (**Supporting Information Table 3**). Four siderophore biosynthetic gene clusters were identified. Genes involved in the biosynthesis of ficellomycin and grincamycin had low similarity, while those involved in the biosynthesis of mirubactin A and bacillibactin were higher in similarity (**Table 1**). Notably, several gene clusters were 100% identical to known genes involved in the production of albaflavenone, informatipeptin, bacillibactin, hopene, geosmin, and ectoine (**Table 1**).

Siderophores detected by LC-MS

Metabolomics was performed on extracts of PD-S100-1 grown in the presence and absence of cerium. Among the features putatively identified are the catecholate siderophore enterobactin, mirubactin A, and three degradation/cyclization products (**Figure 6**). Enterobactin as well as mirubactins B and D were significantly upregulated by cerium elicitation, while mirubactins B and C were downregulated or unchanged.[32]

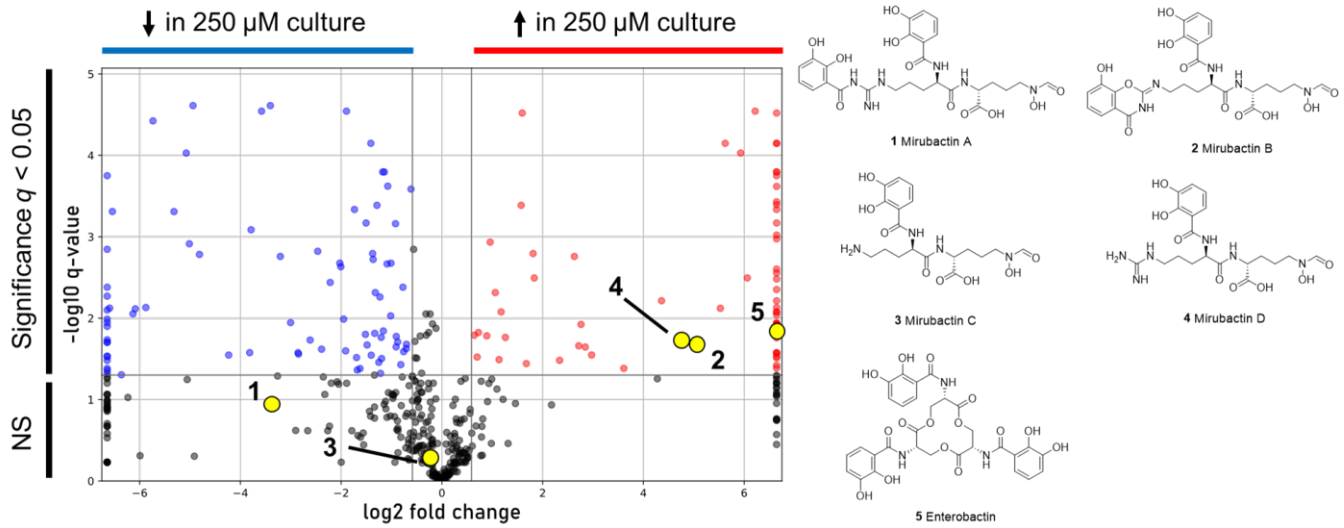


Figure 6. Volcano plot of the LC-MS features in the *Streptomyces* sp. PD-S100-1, with features significantly upregulated in the cerium treatment shown in red, and features significantly downregulated shown in blue. Filled yellow circles indicate features associated with numbered compounds and their structures are shown on the right.

Fragmentation analysis identified the presence of siderophores enterobactin (670.1517 *m/z*, 670.152047 *m/z* calculated for [M+H]) and mirubactin A (605.2199 *m/z*, calculated 605.220734 for [M+H]) (**Figure 7**). The experimental MS/MS fragmentation pattern of mirubactin A contained several peaks characteristic of the compound, including key fragments at 469.2039 *m/z*, 333.1879 *m/z*, and 293.1233 *m/z*. Enterobactin was likewise confirmed based on its MS/MS fragmentation pattern with key peaks at 224.0553 *m/z* and 137.0232 *m/z*. Mirubactin A production was downregulated with cerium treatment, whereas enterobactin production was exclusively produced in the presence of cerium.

Several other catecholates were detected by LC-MS/MS, confirming the presence of siderophores from the enterobactin and mirubactin families (**Table 2**). Further investigation of compounds possessing the characteristic 137.0233 *m/z* dihydroxybenzoyl fragment revealed the presence of the recently reported mirubactin A decomposition products, mirubactins B-D at higher levels in the cerium-elicited culture (**Figure 8**).[32]

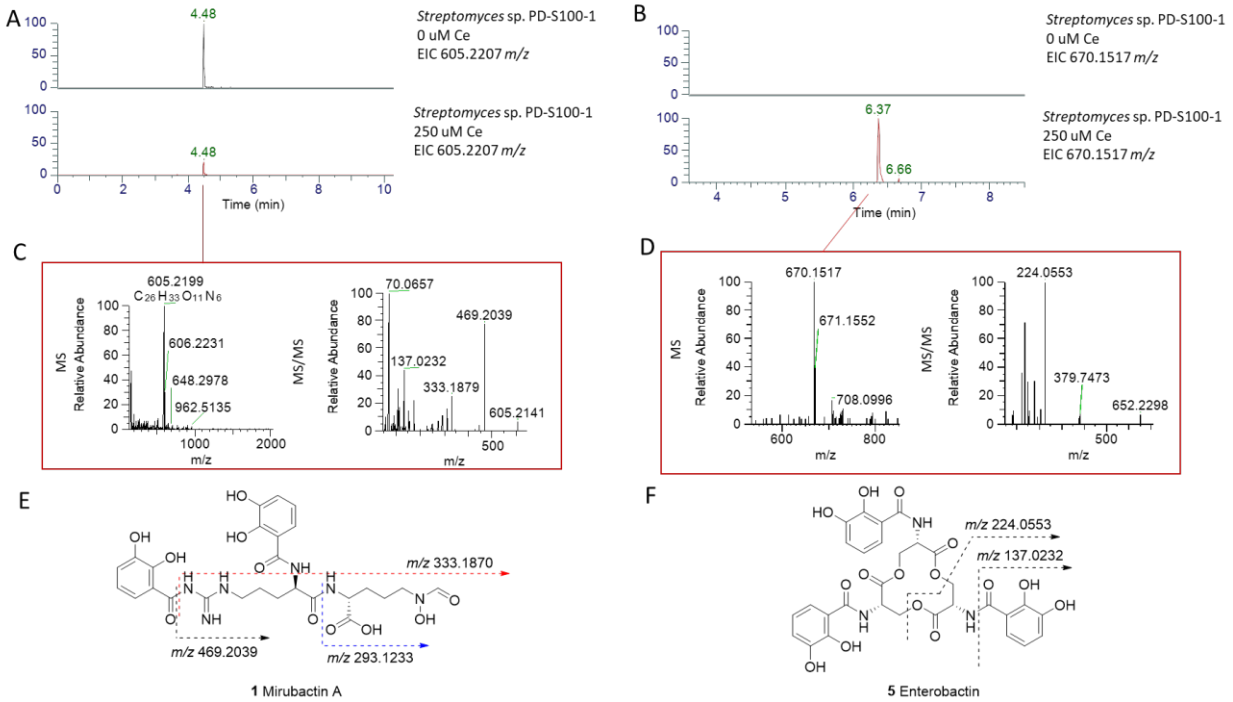


Figure 7. Confirmation of siderophore production by mass spectrometry. Extracted ion chromatograms confirming the production of mirubactin A (left) and enterobactin (right) (A), MS and MS/MS fragmentation (B), and molecular structure with proposed fragmentation pathways (C).

Compound	Retention Time (minutes)	Theoretical [M+H]	Experimental *m/z*
2,3-dihydroxy-*N*-benzoylserine trimer	5.67	688.16261	688.1616
2,3-dihydroxy-*N*-benzoylserine dimer	4.94	465.1145	465.1138
Fimsbactin A	4.70	575.1989	575.1985
Enterobactin	6.37	670.1521	670.1517
Mirubactin A	4.48	605.2207	605.2199
Mirubactin B	4.69	588.1941	588.1935
Mirubactin C	3.25	427.1829	427.1820
Mirubactin D	3.43	469.2036	469.2043

Table 2. Catecholate siderophores confirmed by LC-MS/MS.

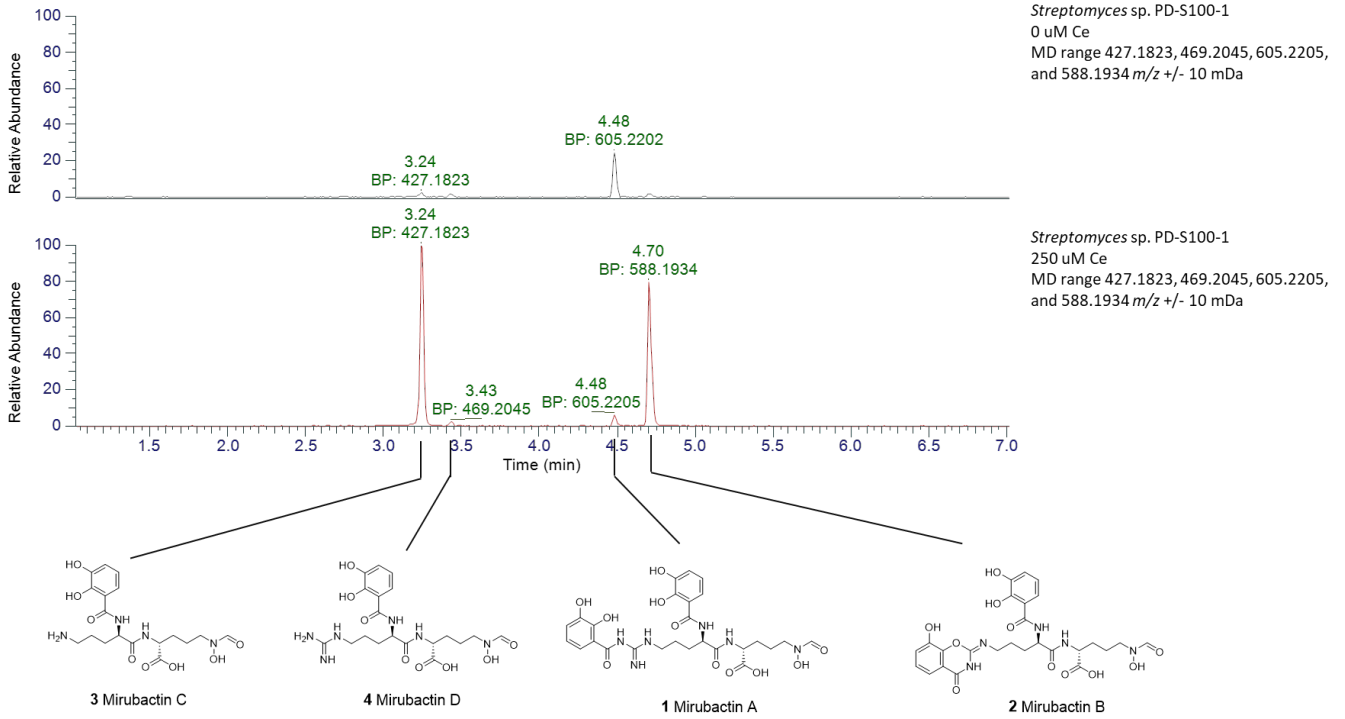


Figure 8. Preliminary extracted ion chromatograms of mirubactin A and its decomposition products in control and cerium-elicited cultures of *Streptomyces sp.* PD-S100-1 showing the marked decrease in mirubactin A in elicited cultures is accompanied by increases in the decomposition products, mirubactins B-D.

DISCUSSION

Phytoremediation involves plants absorbing metals from their roots, oftentimes with the help of bacteria, and accumulating or incorporating them into the plant metabolome.[33] Members of the *Cyperaceae* family have been reported to host siderophore-producing bacteria. Thus, *C. virens* was hypothesized to contain siderophore-producing bacteria in its rhizosphere that may play a role in phytoremediation.[34] Actinobacteria are prolific producers of siderophores; thus, we aimed to identify these bacteria in the *C. virens* rhizosphere.

Sporulating bacteria were preselected from the rhizosphere using a heat treatment as well as selecting for Gram-positive bacteria through the use of NDA and CHX. The PD-S100-1 isolate grew on both GYM-NDA-CHX and marine media plates (**Figure 3**), revealing its salt tolerance.[35] *C. virens* grows in the (sub)tropics of the Americas, where it may inhabit brackish environments, hence its ability for bacteria from its rhizosphere to grow in high salinity media.[8] However, the colonies were fewer in number when the same volume of cells were plated, suggesting that these isolates are salt-tolerant and high-salinity environments decrease growth.

Bacterial siderophore production was confirmed by CAS assays containing a blue iron-chelating dye that turns orange when a molecule outcompetes it for binding iron.[12] While PD-S100-1 displayed no CAS activity on marine agar, the isolate displayed strong density-dependent siderophore production on GYM-NDA-CHX media (**Figure 4**), suggesting the involvement of an autoregulator quorum sensing molecule in the regulation of siderophore production.[36,37] High salinity environments put stress on bacteria, and the larger plant, which may lead to decreased siderophore production; hence, the lack of CAS activity on marine agar.[38] A large orange halo developed around PD-S100-1 in the CAS assay (**Figure 4**), indicating siderophore production. While this is the first report of a siderophore producer within the *C. virens* rhizosphere, siderophore production has been previously reported by the rhizosphere of other strains from the *Cyperus* genus using the CAS assay.[39] However, no significant antimicrobial activity against *Aspergillus niger*, *A. flavus*, or *B. subtilis* was observed, even though Gram-positive bacteria have been reported to be sensitive to extracts from *Cyperus* plants.[40]

Whole genome sequencing revealed PD-S100-1 contained siderophore biosynthetic genes and determined the taxonomic identification of this isolate to be *Streptomyces* sp. PD-S100-1. The high GC content is consistent with actinobacteria, as percentages over 70% often indicate *Streptomyces*.[41] Of the 28 identified biosynthetic gene clusters, only a few gene clusters were similar to those involved in the production of important bioactive molecules. Four siderophore-producing gene clusters were

identified and two were predicted to produce mirubactin A and bacillibactin (**Figure 9**). The other two gene clusters were annotated as siderophore biosynthetic gene clusters based on sequence alignments of experimentally characterized proteins and protein domains but make products unlike any known siderophores, indicating the potential of PD-S100-1 to produce new chemistry. Mirubactin A and bacillibactin biosynthetic gene clusters were higher in similarity to genes known to be involved in their biosynthesis (**Table 1, Figure 9**)[42,43], and both compounds are catecholates consisting of dihydroxybenzoic acid (DHBA) functionalized scaffolds that display strong siderophore activity.[44] Mirubactin A is not only a siderophore, but it also promotes cell wall growth, whereas bacillibactin has antibacterial and antifungal properties.[45,46] The bacillibactin gene cluster may not be expressed under these conditions due to the lack of antimicrobial activity detected under these growth conditions. The identification of these siderophore biosynthetic gene clusters supports the results of the CAS assay and suggests that PD-S100-1 has the potential to produce novel siderophores.

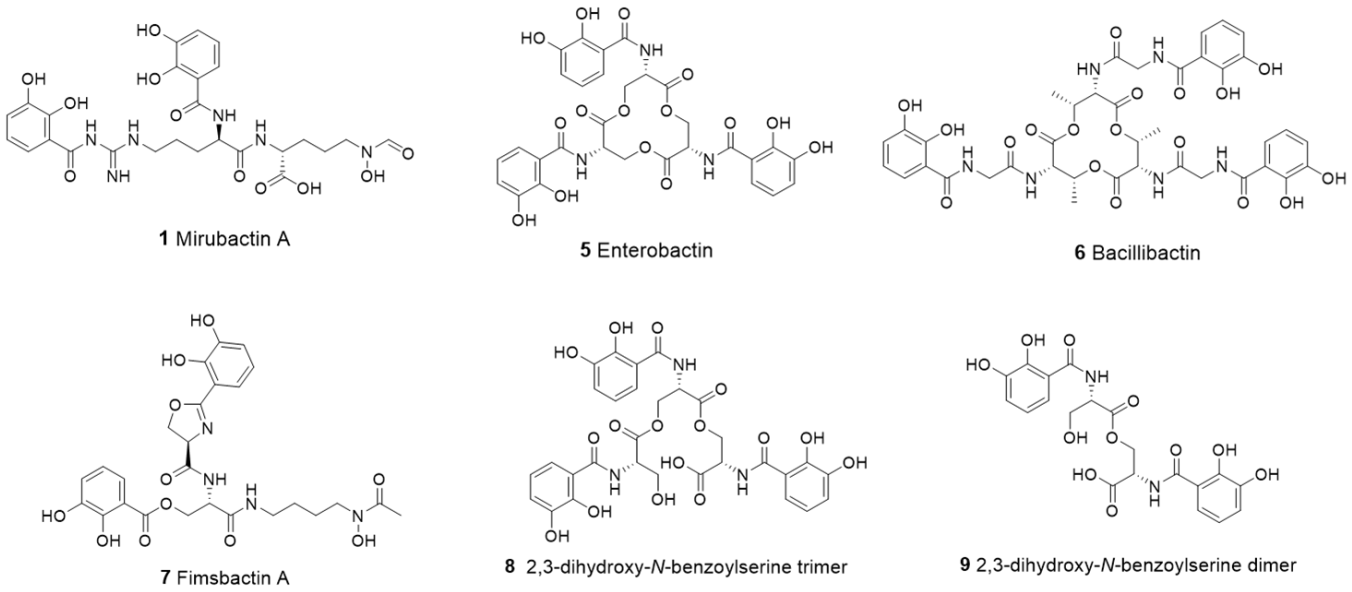


Figure 9. Structures of key catecholate siderophores or families thereof identified in *Streptomyces* sp. PD-S100-1 or its genome using LC-MS/MS.

Several catecholate siderophores were identified by LC-MS (**Table 2**), including the decomposition products of mirubactin A, mirubactins B-D (**Figures 6–9**). The detection of the mirubactins along with their corresponding biosynthetic gene cluster confirmed that this siderophore is produced by *Streptomyces* sp. PD-S100-1, which is consistent with other reports of *Streptomyces* producing mirubactin A.[47] Additionally, the related hybrid siderophore fimsbactin A was detected (**Figure 9**). Most secondary metabolite production was upregulated (**Figure 7**) when *Streptomyces* sp. PD-S100-1 was grown in the presence of cerium, a rare earth element (REE) that can trigger the production of secondary metabolites from biosynthetic genes not always expressed (i.e., cryptic metabolites).[11] Interestingly, Khan *et al.* found that multiple members of the *Cyperus* genus accumulate or tolerate REEs, which may also occur when PD-S100-1 is grown in the presence of *C. virens* and cerium.[48] The addition of low concentrations of REEs has been shown to have beneficial effects on the growth of many plants, including increasing their iron uptake.[48]

The observed decrease in mirubactin A in cultures treated with cerium was unexpected, as significant increases in enterobactin were detected and both of these molecules are catecholates (**Figure 6**). Furthermore, increases in mirubactins B-D, known decomposition products of mirubactin A,[32] were also observed, suggesting that cerium may not decrease the production of mirubactins but rather accelerate their decomposition to a diverse set of related hydroxamate siderophores. It is unclear if this finding has practical implications for the metal-chelating phenotype of *Streptomyces sp.* PD-S100-1. However, mirubactins C-D were reported to exhibit reduced iron complexing capability and the unusual cyclic amide moiety in mirubactin B may modulate its metal chelation profile.[32]

Enterobactin was detected in extracts of PD-S100-1 (**Figure 6–7**), which has also been reported in two other *Streptomyces* species.[49] Although enterobactin was produced, it was challenging to identify its biosynthetic gene cluster. While the NRPS modules in the annotated bacillibactin gene cluster putatively code only for serine and dihydrobenzoic acid (DHBA)—which is congruent with enterobactin rather than bacillibactin—the architecture of this gene cluster differed from other reported enterobactin gene clusters. While bacillibactin was not detected in extracts, the enterobactin gene cluster may be undetectable because it could be split between contigs or some functions of the biosynthetic gene cluster (i.e., shikimate synthesis) are provided by the mirubactin

biosynthetic machinery, rendering it as nonfunctional in the absence of another catecholate siderophore gene cluster. Enterobactin increased in the cerium treatment, following the overall trend of cerium positively impacting secondary metabolite regulation. Interestingly, mirubactin A was downregulated, whereas enterobactin was upregulated, as these belong to the same family of siderophores. It is unclear why there wasn't a similar increase in enterobactin's scaffold and DHBA hydrolysis like mirubactin A, though the cyclic scaffold of enterobactin may increase its stability relative to the linear mirubactin scaffold. Despite this, low levels of the DHBA-serine dimer and trimer, which may be hydrolysis products, were detected at low levels (**Table 2, Figure 9**).

PD-S100-1 not only produces catecholate siderophores but also has the genetic potential to make several other environmentally relevant secondary metabolites that were undetected under these growth conditions and LC/MS methods. For example, biosynthetic gene clusters dedicated to making the osmoregulator ectoine were identified. Ectoine can stabilize microorganisms under high osmotic stress conditions and prevent desiccation.[50] PD-S100-1 grew on the marine agar (**Figure 3**), and the ectoine gene cluster may enable it to overcome the environmental stress of the high salinity. Furthermore, other gene clusters were annotated as being involved in the production of the antimicrobials albaflavenone and informatipeptin.[51,52] While PD-S100-1 did not exhibit antibacterial or antifungal properties against *B. subtilis*, *A. niger*, and *A. flavus* under the growth conditions used, this gene cluster suggests possible antimicrobial activity of bacteria in the *C. virens* rhizosphere.

CONCLUSIONS

Streptomyces sp. PD-S100-1 was isolated from the *C. virens* rhizosphere and found to produce siderophores in CAS plate assays. The genome of this isolate was sequenced and found to contain biosynthetic gene clusters for antimicrobials, such as albaflavenone and informatipeptin, and siderophore production via mirubactin and bacillibactin, supporting the results from the CAS plate assays. Metabolomics connected genes to the actual molecules produced by *Streptomyces* sp. PD-S100-1, confirming the production of mirubactin A and enterobactin, among other catecholate members of their families. *Streptomyces* sp. PD-S100-1 may play useful roles in iron acquisition in *C. virens*. Importantly, linking biosynthetic genes to their producing metabolites provides opportunities to source metabolites, such as siderophores, and perform genetic manipulation to modify metabolite production, especially in the context of *Cyperus* phytoremediation.

ACKNOWLEDGMENTS

This work was supported by the Faculty Early Career Development (CAREER) Award from the National Science Foundation (NSF) (grant number 2236643 for P.D., R.S, R.R., and L-A.G.). We are also extremely grateful to Sophie Collet for assisting with the formatting of this manuscript.

Supporting information available here.

REFERENCES

1. Vijay, K., Shibasini, M., Sivasakthivelan, P., and Kavitha, T. (2023) Microbial siderophores as molecular shuttles for metal cations: sources, sinks and application perspectives, *Arch Microbiol* 205, 322. *https://doi.org/10.1007/s00203-023-03644-3*
2. Xie, B., Wei, X., Wan, C., Zhao, W., Song, R., Xin, S., and Song, K. (2024) Exploring the Biological Pathways of Siderophores and Their Multidisciplinary Applications: A Comprehensive Review, *Molecules* 29, 2318. *https://doi.org/10.3390/molecules29102318*
3. Wang, W., Qiu, Z., Tan, H., and Cao, L. (2014) Siderophore production by actinobacteria, *Biometals* 27, 623–631. *https://doi.org/10.1007/s10534-014-9739-2*
4. Basumatary, B., Bordoloi, S., and Sarma, H.P. (2012) Crude Oil-Contaminated Soil Phytoremediation by Using *Cyperus brevifolius* (Rottb.) Hassk, *Water Air Soil Pollut* 223, 3373–3383. *https://doi.org/10.1007/s11270-012-1116-6*
5. Senanu, L.D., Kranjac-Berisavljevic, G., and Cobbina, S.J. (2023) The use of local materials to remove heavy metals for household-scale drinking water treatment: A review, *Environmental Technology & Innovation* 29, 103005. *https://doi.org/10.1016/j.eti.2023.103005*
6. Chandra, R., and Yadav, S. (2011) Phytoremediation of Cd, Cr, Cu, Mn, Fe, Ni, Pb and Zn from Aqueous Solution Using Phragmites Cummunis, Typha Angustifolia and *Cyperus Esculentus*, *International Journal of Phytoremediation* 13, 580–591. *https://doi.org/10.1080/15226514.2010.495258*
7. de Souza, D.M., da Silva, J. de L., Ludwig, L. da C., Petersen, B.C., Brehm, F.A., Modolo, R.C.E., De Marchi, T.C., Figueiredo, R., and Moraes, C.A.M. (2023) Study of the phytoremediation potential of native plant species identified in an area contaminated by volatile organic compounds: a systematic review, *Int J Phytoremediation* 25, 1524–1541. *https://doi.org/10.1080/15226514.2023.2170974*
8. Tucker, G.C., Marcks, B.G., and Carter, J.R. *Cyperus virens* - FNA, *Flora North America*. http://floranorthamerica.org/Cyperus_virens
9. *Cyperus virens* Michx *USDA Plants Database*. https://plants.usda.gov/home/plantProfile?symbol=CYVI2

10. *Cyperus virens* Michx. *Royal Botanical Gardens Kew Plants of the Online World.* http://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:306303-1

11. Tanaka, Y., Hosaka, T., and Ochi, K. (2010) Rare earth elements activate the secondary metabolite–biosynthetic gene clusters in *Streptomyces coelicolor* A3(2), *J Antibiot* 63, 477–481. *https://doi.org/10.1038/ja.2010.53*

12. Louden, B.C., Haarmann, D., and Lynne, A.M. (2011) Use of Blue Agar CAS Assay for Siderophore Detection, *Journal of Microbiology & Biology Education. https://doi.org/10.1128/jmbe.v12i1.249*

13. Goecks, J., Nekrutenko, A., Taylor, J., and The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, *Genome Biology* 11, R86. *https://doi.org/10.1186/gb-2010-11-8-r86*

14. Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics* 30, 2114–2120. *https://doi.org/10.1093/bioinformatics/btu170*

15. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

16. Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2016) Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. Preprint at bioRxiv, *https://doi.org/10.1101/096412 https://doi.org/10.1101/096412*

17. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018) Versatile genome assembly evaluation with QUAST-LG, *Bioinformatics* 34, i142–i150. *https://doi.org/10.1093/bioinformatics/bty266*

18. Mikheenko, A., Valin, G., Prjibelski, A., Saveliev, V., and Gurevich, A. (2016) Icarus: visualizer for de novo assembly evaluation, *Bioinformatics* 32, 3321–3323. *https://doi.org/10.1093/bioinformatics/btw379*

19. Mikheenko, A., Saveliev, V., and Gurevich, A. (2016) MetaQUAST: evaluation of metagenome assemblies, *Bioinformatics* 32, 1088–1090. *https://doi.org/10.1093/bioinformatics/btv697*

20. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics* 29, 1072–1075. *https://doi.org/10.1093/bioinformatics/btt086*

21. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research* 25, 3389–3402. *https://doi.org/10.1093/nar/25.17.3389*

22. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009) BLAST+: architecture and applications, *BMC Bioinformatics* 10, 421. *https://doi.org/10.1186/1471-2105-10-421*

23. Cock, P.J.A., Chilton, J.M., Grüning, B., Johnson, J.E., and Soranzo, N. (2015) NCBI BLAST+ integrated into Galaxy, *GigaScience* 4, s13742-015-0080–0087. *https://doi.org/10.1186/s13742-015-0080-7*

24. Sayers, E.W., Beck, J., Bolton, E.E., Brister, J.R., Chan, J., Connor, R., Feldgarden, M., Fine, A.M., Funk, K., Hoffman, J., *et al.* (2024) Database resources of the National Center for Biotechnology Information in 2025, *Nucleic Acids Res* 53, D20–D29. *https://doi.org/10.1093/nar/gkae979*

25. Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., *et al.* (2015) RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes, *Sci Rep* 5, 8365. *https://doi.org/10.1038/srep08365*

26. Cuccuru, G., Orsini, M., Pinna, A., Sbardellati, A., Soranzo, N., Travaglione, A., Uva, P., Zanetti, G., and Fotia, G. (2014) Orione, a web-based framework for NGS analysis in microbiology, *Bioinformatics* 30, 1928–1929. *https://doi.org/10.1093/bioinformatics/btu135*

27. Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation, *Bioinformatics* 30, 2068–2069. *https://doi.org/10.1093/bioinformatics/btu153*

28. Blin, K., Shaw, S., Augustijn, H.E., Reitz, Z.L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B.R., Metcalf, W.W., Helfrich, E.J.N., *et al.* (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation, *Nucleic Acids Research* 51, W46–W50. *https://doi.org/10.1093/nar/gkad344*

29. Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., and Arita, M. (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, *Nat Methods* 12, 523–526. *https://doi.org/10.1038/nmeth.3393*

30. Samples, R.M., Puckett, S.P., and Balunas, M.J. (2023) Metabolomics Peak Analysis Computational Tool (MPACT): An Advanced Informatics Tool for Metabolomics and Data Visualization of Molecules from Complex Biological Samples, *Anal Chem* 95, 8770–8779. *https://doi.org/10.1021/acs.analchem.2c04632*

31. Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., *et al.* (2016) Sharing and community curation of mass spectrometry data with GNPS, *Nat Biotechnol* 34, 828–837. *https://doi.org/10.1038/nbt.3597*

32. Pu, H., Jiang, T., Peng, D., Xia, J., Gao, J., Wang, Y., Yan, X., Huang, X., Duan, Y., and Huang, Y. (2022) Degradation of mirubactin to multiple siderophores with varying Fe(III) chelation properties, *Org Biomol Chem* 20, 5066–5070. *https://doi.org/10.1039/D2OB00942K*

33. Montreemuk, J., Stewart, T.N., and Prapagdee, B. (2024) Bacterial-assisted phytoremediation of heavy metals: Concepts, current knowledge, and future directions, *Environmental Technology & Innovation* 33, 103488. *https://doi.org/10.1016/j.eti.2023.103488*

34. Wang, S., Wang, J., Zhou, Y., Huang, Y., and Tang, X. (2022) Prospecting the plant growth–promoting activities of endophytic bacteria Franconibacter sp. YSD YN2 isolated from *Cyperus esculentus* L. var. sativus leaves, *Annals of Microbiology* 72, 1. *https://doi.org/10.1186/s13213-021-01656-2*

35. PanReac AppliChem Marine Agar (Dehydrated Culture Media) for microbiology,

36. Heffernan, J.R., Wildenthal, J.A., Tran, H., Katumba, G.L., McCoy, W.H., and Henderson, J.P. (2024) Yersiniabactin is a quorum-sensing autoinducer and siderophore in uropathogenic *Escherichia coli*, *mBio* 15, e0027723. *https://doi.org/10.1128/mbio.00277-23*

37. Khan, M., Bhargava, P., and Goel, R. (2019) Quorum Sensing Molecules of Rhizobacteria: A Trigger for Developing Systemic Resistance in Plants. In Plant Growth Promoting Rhizobacteria for Sustainable Stress Management : Volume 1: Rhizobacteria in Abiotic Stress Management, R. Z. Sayyed, N. K. Arora, and M. S. Reddy, eds. (Springer), pp. 117–138. *https://doi.org/10.1007/978-981-13-6536-2_7*

38. Sultana, S., Alam, S., and Karim, M.M. (2021) Screening of siderophore-producing salt-tolerant rhizobacteria suitable for supporting plant growth in saline soils with iron limitation, *Journal of Agriculture and Food Research* 4, 100150. *https://doi.org/10.1016/j.jafr.2021.100150*

39. C. Sarathambal, P.J. Khankhane, and D.K. Singh (2015) Characterization of metal tolerant and plant growth promoting rhizobacteria colonizing in weed rhizosphere, *Weed Science for Sustainable Agriculture, Environment and Biodiversity* 1, 4.

40. Taheri, Y., Herrera-Bravo, J., Huala, L., Salazar, L.A., Sharifi-Rad, J., Akram, M., Shahzad, K., Melgar-Lalanne, G., Baghalpour, N., Tamimi, K., *et al.* (2021) *Cyperus spp.*: A Review on Phytochemical Composition, Biological Activity, and Health-Promoting Effects, *Oxidative Medicine and Cellular Longevity* 2021, 4014867. *https://doi.org/10.1155/2021/4014867*

41. Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G.F., Chater, K.F., and Sinderen, D. van (2007) Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum, *Microbiology and Molecular Biology Reviews : MMBR* 71, 495. *https://doi.org/10.1128/MMBR.00005-07*

42. Giessen, T.W., Franke, K.B., Knappe, T.A., Kraas, F.I., Bosello, M., Xie, X., Linne, U., and Marahiel, M.A. (2012) Isolation, Structure Elucidation, and Biosynthesis of an Unusual Hydroxamic Acid Ester-Containing Siderophore from *Actinosynnema mirum, J Nat Prod* 75, 905–914. *https://doi.org/10.1021/np300046k*

43. Chakraborty, K., Kizhakkekalam, V.K., Joy, M., and Chakraborty, R.D. (2022) Bacillibactin class of siderophore antibiotics from a marine symbiotic *Bacillus* as promising antibacterial agents, *Appl Microbiol Biotechnol* 106, 329–340. *https://doi.org/10.1007/s00253-021-11632-0*

44. Timofeeva, A.M., Galyamova, M.R., and Sedykh, S.E. (2022) Bacterial Siderophores: Classification, Biosynthesis, Perspectives of Use in Agriculture, *Plants* 11, 3065. *https://doi.org/10.3390/plants11223065*

45. Kepplinger, B., Wen, X., Tyler, A.R., Kim, B.-Y., Brown, J., Banks, P., Dashti, Y., Mackenzie, E.S., Wills, C., Kawai, Y., *et al.* (2022) Mirubactin C rescues the lethal effect of cell wall biosynthesis mutations in *Bacillus subtilis*, *Front Microbiol* 13. *https://doi.org/10.3389/fmicb.2022.1004737*

46. Dimopoulou, A., Theologidis, I., Benaki, D., Koukounia, M., Zervakou, A., Tzima, A., Diallinas, G., Hatzinikolaou, D.G., and Skandalis, N. (2021) Direct Antibiotic Activity of Bacillibactin Broadens the Biocontrol Range of *Bacillus amyloliquefaciens* MBI600, *mSphere* 6, e0037621. *https://doi.org/10.1128/mSphere.00376-21*

47. Agathokleous, E., Zhou, B., Geng, C., Xu, J., Saitanis, C.J., Feng, Z., Tack, F.M.G., and Rinklebe, J. (2022) Mechanisms of cerium-induced stress in plants: A meta-analysis, *Science of The Total Environment* 852, 158352. *https://doi.org/10.1016/j.scitotenv.2022.158352*

48. Khan, A.M. (2017) Accumulation, uptake and bioavailablility of rare earth elements (REES) in soil grown plants from ex-mining area in Perak, Malaysia, *Appl Ecol Env Res* 15, 117–133. *https://doi.org/10.15666/aeer/1503_117133*

49. Fiedler, H.-P., Krastel, P., Müller, J., Gebhardt, K., and Zeeck, A. (2001) Enterobactin: the characteristic catecholate siderophore of Enterobacteriaceae is produced by *Streptomyces* species, *FEMS Microbiology Letters* 196, 147–151. *https://doi.org/10.1111/j.1574-6968.2001.tb10556.x*

50. Reshetnikov, A.S., Khmelenina, V.N., Mustakhimov, I.I., and Trotsenko, Y.A. (2011) Genes and Enzymes of Ectoine Biosynthesis in Halotolerant Methanotrophs. In Methods in Enzymology (Elsevier), pp. 15–30. *https://doi.org/10.1016/B978-0-12-386905-0.00002-4*

51. Zhao, B., Lin, X., Lei, L., Lamb, D.C., Kelly, S.L., Waterman, M.R., and Cane, D.E. (2008) Biosynthesis of the Sesquiterpene Antibiotic Albaflavenone in *Streptomyces coelicolor* A3(2) *, *Journal of Biological Chemistry* 283, 8183–8189. *https://doi.org/10.1074/jbc.M710421200*

52. Deflandre, B., Stulanovic, N., Planckaert, S., Anderssen, S., Bonometti, B., Karim, L., Coppieters, W., Devreese, B., and Rigali, S. (2022) The virulome of *Streptomyces scabiei* in response to cello-oligosaccharide elicitors, *Microbial Genomics* 8, 000760. *https://doi.org/10.1099/mgen.0.000760*

ABOUT STUDENT AUTHOR

Phoebe Dennison is currently a chemistry and Spanish double major at Smith College. She completed this research for a chemical ecology course (CHM 350; Chemical Ecology: Catch Fe If You Can) but continues to conduct research in other areas of chemistry. She currently does surface chemistry research with Professor Kate Queeney in the Department of Chemistry. Phoebe will graduate in the spring of 2026 and plans to pursue a career in human health.

PRESS SUMMARY

Plants and root-associated microbes can be used to clean up heavy metal-contaminated environments using natural metal-binding molecules or siderophores. Some *Cyperus* plants have been reported to bind iron in wastewater, making them useful for cleaning up metal pollution. Herein, we examined whether siderophore-producing bacteria in the rhizosphere of *Cyperus virens*, an understudied member of this genus, could be used to treat metal pollution. Using an assay that changes color in the presence of siderophores, we isolated siderophore-producing bacteria. By analyzing their DNA, we discovered one of the bacteria was *Streptomyces* sp. PD-100-01 and contained genes that could make bioactive compounds, including siderophores. Furthermore, chemical analyses of *Streptomyces* sp. PD-100-01 cultures detected several catecholate siderophores, including mirubactin A and enterobactin. Together, these data suggest *C. virens* and its root-associated bacteria may be valuable tools for cleaning up metal pollution.

Utilizing a Low-Cost Telescope to Produce Color-Magnitude Diagrams of an Open Cluster

Clay Reece & Inseok Song*

Department of Physics and Astronomy, University of Georgia, Athens, GA

https://doi.org/10.33697/ajur.2025.159

*Student: clayreece@uga.edu, clayburkr@gmail.com**
Mentor: song@uga.edu

ABSTRACT
Hands-on learning is critical for students' education, especially for more complex subjects, such as astrophysics. Astrophysics students are often taught more theoretical concepts than practical research skills, and thus have trouble adjusting to learning methods that produce new astronomical discoveries. This is a complicated problem, as it is difficult to obtain astronomical equipment due to high costs; large telescopes and good camera equipment are expensive. Many schools simply do not have the budget to spend on costly telescopes and cameras. This is an enormous drawback for students' education, as they cannot apply the knowledge they gain in the classroom setting to further their understanding of the astrophysics topics they learn. This paper aims to show that producing scientific data is possible even with low-cost, entry-level astronomical equipment, which can be utilized in classroom settings to help teach students about typical astronomical research methodology through hands-on learning. Many tools exist in astronomy that are either open source or relatively cheap, and this should not go unnoticed by students and educators with a low budget. This paper aims to utilize these tools to show that astronomical data can be retrieved with a great deal of success.

KEYWORDS
Astronomy Education; Hands-On Learning; Education; Accessibility; Low-Cost; Telescope; Photometry; Color-Magnitude Diagram; Open Cluster; Pleiades

INTRODUCTION

Hands-on learning has enormously benefited students' understanding of certain topics.[1] It allows students to both physically do activities and apply the knowledge they have gained, helping them achieve a stronger understanding of the topic at hand.[2] Hands-on methodology is critical for the education and engagement of students worldwide, and Kanwal et al. have shown that inexpensive learning material improves students' ability to learn more through hands-on methods compared to hands-off methods.[3]

Photometry, the measurement of stars' brightnesses, is one of the most important methods of data retrieval in observational stellar astronomy research. Performing photometry can indicate much about the formation and evolution of stars, which can give us a stronger understanding of how our sun will evolve and how it has changed since its initial formation. Performing photometry in different light wavelengths and comparing them can provide strong indications of where a star is along its life and what type of star it is. A common way for astronomers to compare the relative wavelengths of stars is to create a color-magnitude diagram, a graph comparing stars' magnitude in a certain range of wavelengths to a color index, which is the difference between smaller and larger wavelength ranges' magnitudes. This project used red, green, and blue image color channels to generate color indices. The reasoning behind making a color-magnitude diagram is that it strongly correlates with a Hertzsprung-Russell (HR) diagram, an example of which can be seen in **Figure 1**. This graph compares stars' luminosities to their surface temperature, or in some cases, due to the strong correlation, the color index. This correlation implies that a certain relationship exists between the intrinsic brightness and temperature of a star. A more massive star is hotter and brighter than a lower-mass star. Using HR diagrams, astronomers can classify stars by their spectral class, age, and mass, among many other things.

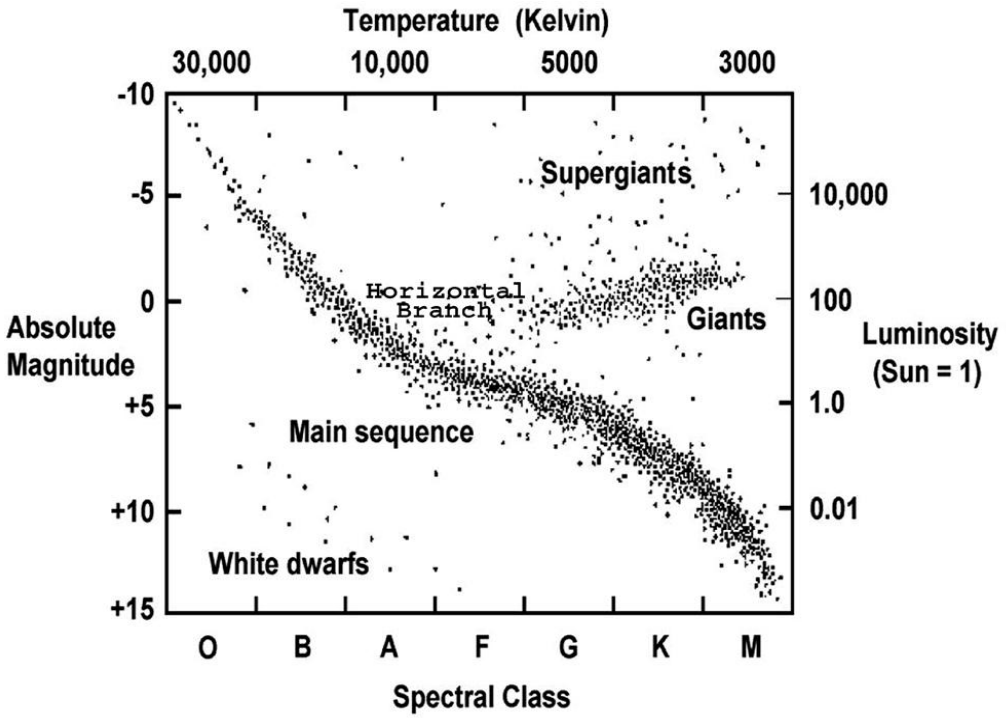


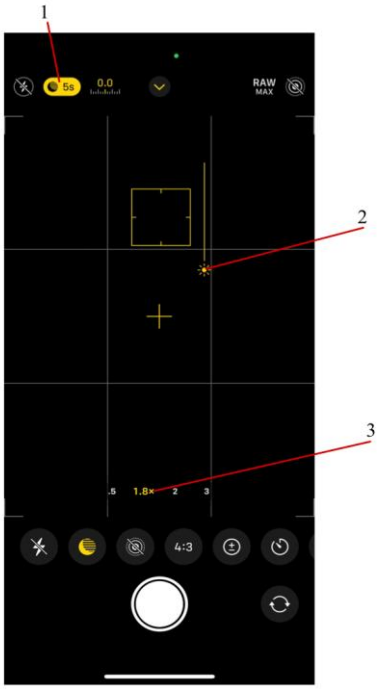
Figure 1. HR diagram with temperature and spectral class on the x-axis, and absolute magnitude and solar luminosity on the y-axis. Image credit goes to the Harvard-Smithsonian Center for Astrophysics.[4]

This project uses a low-cost telescope to generate a color-magnitude diagram for the Pleiades (also known as Messier 45, or M45) open cluster. It is the closest (400 light years away), young (125 million years old), and rich open cluster, with 2000 known members. The Pleiades were chosen due to their high brightness and relatively young age of about 100 million years, meaning that a larger majority of the bright stars are still within the main sequence, allowing a more obvious trend to be seen in the color-magnitude diagrams.[5] The goal of this project is to show that it is possible to use low-cost equipment in order to teach students about the processes that go into observational astronomy research and produce results that are generated with data gathered by the students themselves. This project will encourage the idea that astronomy is accessible to students and teachers with a relatively low budget.

METHODS AND PROCEDURES

Several steps were involved in this project. The primary objectives were observing, image preprocessing, performing the photometry, and generating a color-magnitude diagram from the magnitudes.

Observations were performed at Iron Horse in Watkinsville, Georgia. This location was chosen due to its relatively short distance from the University of Georgia campus and relatively low light pollution, with a radiance value of 0.90nW/cm^2 as of 2023, and no information regarding recent anti-light pollution efforts since then could be found, so it likely has not decreased much more than this if at all.[6] The telescope used was the Celestron FirstScope with an aperture of 76 mm and a focal length of 300 mm, and photographs were taken using the 20 mm eyepiece that comes with the telescope. The camera used for imaging M45 was the iPhone 15 Pro, using its 24 mm lens and the native Apple camera app, as shown in **Figure 2**. The iPhone 15 Pro's camera has a 48-megapixel complementary metal oxide semiconductor (CMOS) sensor that utilizes a quad Bayer color filter array (CFA). The typical telescope and camera setup can be seen in **Figure 3**. Camera exposure was set to a minimum to minimize the International Organization of Standardization (ISO) sensitivity, and exposure times ranged from 1/16 second to 1.1 seconds using 1.8x digital magnification, and the images were taken with the ProRAW camera mode. Everything else was left as the default option. Observations were done over five nights, and the images taken were put into an observation log, as shown in **Figure 4**. When performing a saturation check on the obtained target images, the stars in the images have relatively consistent saturations with brightness values typically well below 100, implying that they are not oversaturated.



Key Features

1. "Night Mode" – exposure time control, 0-10 second exposure time range
2. Exposure – set to a minimum to minimize ISO
3. Magnification – Set to 1.8x for all gathered images

Figure 2. The iPhone 15 Pro's camera app interface.

Figure 3. The telescope setup consists of the Celestron FirstScope, an iPhone 15 Pro, a Celestron phone mount, and a folding table.

Date	EST Time Range (HH:MM)	Image Types	Exposure Time Range (s)	Usability (good/bad
9/8/2024	21:45-22:50	Tests	1.1-10.0	25 test
9/20/2024	21:23-21:33	Tests	1.1	60 test
10/10/2024	05:34-05:52	Science, Bias, Dark	1/16-1.1	48/3 science, 0/10 bias, 5/0 dark
10/23/2024	00:59-01:00	Bias	1/18-1.1	0/10 bias
11/12/2024	00:50	Bias	1/16	10/0 bias

Figure 4. The observation log. This includes the date, the 24-hour time range in which images were gathered, what types of images were gathered, the ranges of exposure times for the images, and how many of each type of image were usable for this project.

Preprocessing in this project involved three major steps: stacking, bias subtraction, and dark correction. Stacking is the process of adding each spatially calibrated image array to obtain larger brightness values for stars. Bias subtraction removes any imperfections that are caused by defections in CMOS cameras and helps to reduce noise in the images. Dark correction removes unwanted signals caused by hot pixels and electronics inside the camera by subtracting a median-combined dark image. The observations yielded 5 dark images, 10 bias images, and 51 target images. The preprocessing was done using DeepSkyStacker, which contains algorithms that find stars in images to rotate, adjust, and stack them and perform the proper preprocessing (calibration, bias subtraction, dark subtraction, and bad pixel masking). Users must import the proper unprocessed target images, such as the one in **Figure 5,** as well as calibration frames. These calibration frames being bias images, gathered by covering your camera with something dark, such as the lens cap, and taking zero (or as close to as possible) second-exposure images, dark images, gathered by repeating the setup for bias images, but choosing the same exposure time and environment as the data images that you are gathering, and flat frames, gathered by imaging a uniformly lit surface, such as a white blanket or a clear sky at twilight. This project did not use flat calibration. DeepSkyStacker uses these input images to create a single final preprocessed image that can then be exported as a Flexible Image Transport System (FITS) file, such as the one in **Figure 6**. This final image had an effective exposure time of 13 seconds.

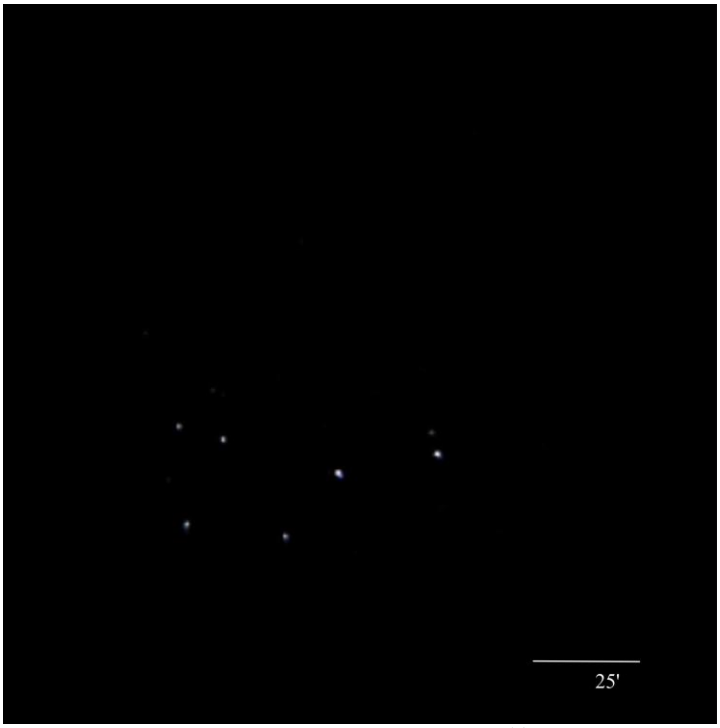


Figure 5. An unprocessed DNG image of M45 cropped and edited to make the stars more visible. This editing was not done for the actual data. The total field of view is approximately two degrees.

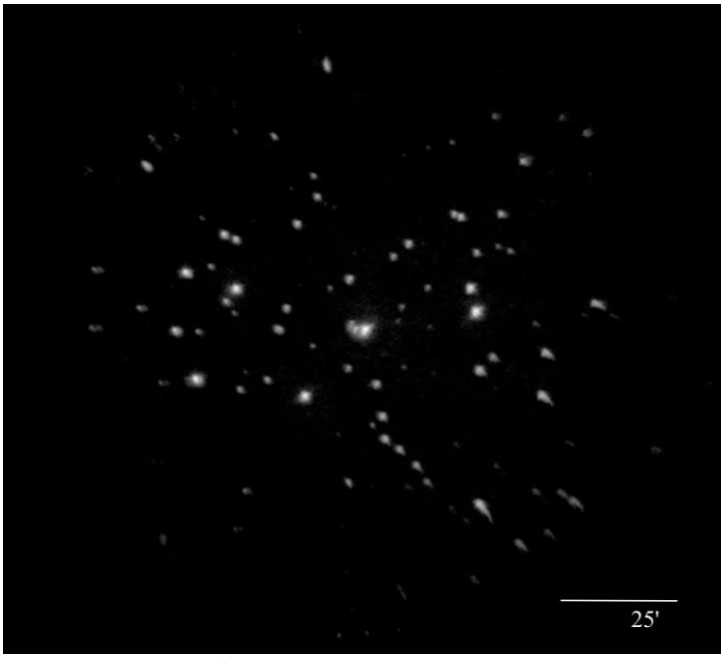


Figure 6. The stacked and preprocessed image of M45.

Once the preprocessed image was obtained, analysis was conducted using a Python script to perform the photometry. Python was chosen due to its large number of statistical and astronomical libraries. The four used in this project were NumPy, photutils, Astropy, and matplotlib. Using photutils' Image Reduction and Analysis Facility (IRAF) star finder, it was possible to locate stars in the image and perform aperture photometry on the located stars, as shown in **Figure 7.** To detect stars, the settings for IRAF star finder were a full-width half maximum (FWHM) of 5.0 pixels and a threshold of 5×10^{-5}. The FWHM is a measure of how sharp or blurry a star looks in an image, and it is measured as the pixel width of a typical stellar image at half of its peak brightness. The threshold describes the minimum possible value for pixel brightness that the IRAF star finder would use to distinguish stars from background noise. The rest of the inputs were left as default. The sources found with IRAF star finder were from the red image channel; there seemed to be some background noise in both the green and blue channels that interfered with star detection in the blue and green channels. The radii used for the aperture were 20 pixels, and the inner and outer radii for the annuli were 28 and 38 pixels, respectively. These values were chosen because they appeared to minimize the statistical error in the resultant data while maintaining a large enough set of apertures to contain every star and also have enough background subtraction not to include other sources for the most part. The result of the photometry was a flux that could be converted to a magnitude by using **Equation 1,** where M is the magnitude, f is the flux, and t is the exposure time. In the script, the zero point was chosen to be zero, as this project focused on performing relative photometry rather than absolute photometry.

$$M = zeropoint - 2.5log_{10}(f/t) \quad \textbf{Equation 1.}$$

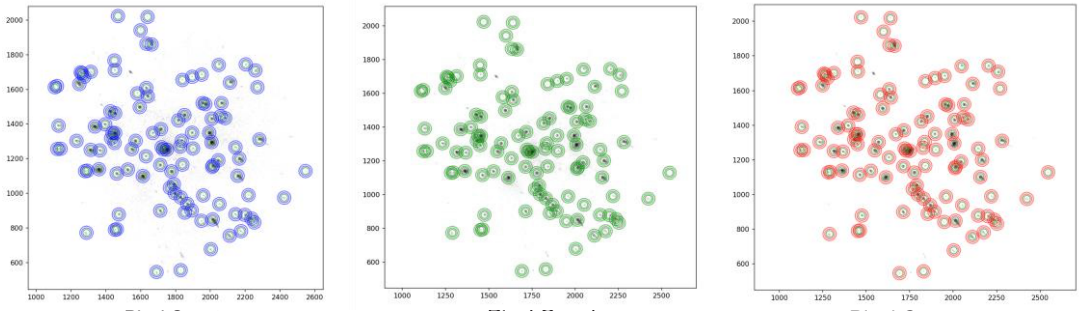


Figure 7. The images with the stars located and the apertures & annuli around them.

After the photometry was performed and errors were calculated, creating a color-magnitude diagram was relatively simple. Using matplotlib's subplot function, the individual magnitudes of each star in red, green, and blue on the y-axis and the difference between the different magnitudes were plotted.

RESULTS

The result of this project is the set of color-magnitude diagrams for M45 shown in **Figure 8**, created from data gathered with a low-cost telescope. Since the errors involved with these images are unquantifiable with our methodology, we opted to take the standard error of the flux for each data point using **Equation 2**, where N is the count of data points, in this case, being pixels within the apertures. From there, the errors were propagated to convert these errors to magnitude errors for each star using **Equation 3**, where dM is the magnitude error, f is the star's flux, and df is the flux error.

$$Standard\ Error\ = (Standard\ Deviation)/\sqrt{N}$$

Equation 2.

$$dM = (2.5/f) * df$$

Equation 3.

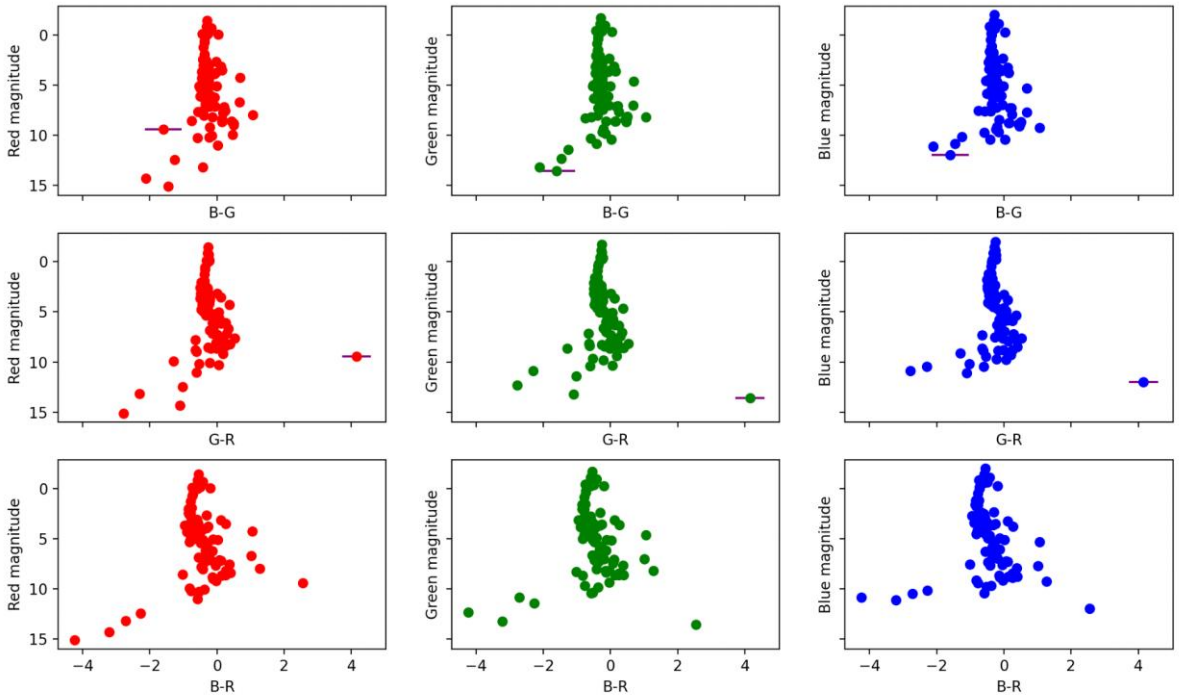


Figure 8. The resulting color-magnitude diagrams. The colors of the plot symbols correspond to the specific band (R, G, B band) plotted in the vertical axis in each column of subplots.

Not including personal costs, such as the iPhone 15 Pro and the computers used to write and run the code, the only necessary expenses were the telescope itself, and the mount used to place the phone on the telescope for more stable data. The total cost of this project was just under $100, not including personal equipment, such as the phone, or computers involved.

DISCUSSION

The maximum statistical errors for the red, green, and blue magnitudes are 0.055, 0.037, and 0.207, respectively, which include all data points. When removing data points that are dimmer than the maximum magnitude that a telescope can theoretically detect, also known as the "limiting magnitude," the maximum red and green magnitude statistical errors decrease slightly, being 0.055 and 0.031, respectively, but the maximum blue magnitude statistical error decreases to 0.047. These errors are low enough for strong confidence in the data, especially when investigating limiting magnitude-corrected data points, as they are four orders of magnitude below the range of the data. Despite the low error, there are some potential reasons for the larger errors that appear in some data points. One of which is the annuli and contamination from other stars. Some stars' annuli contain another star within them. In these cases, when the background is subtracted, it is subtracting another star's flux entirely, which may drastically affect the true value of the flux. This can be adjusted by changing the radius size of the annuli, but finding a sweet spot for annulus size

can be difficult. For this research, the code was run with several annulus sizes, and a radius of 20 pixels minimized the data's maximum statistical errors.

These color-magnitude diagrams show a decently strong trend in line with the main sequence, which can be seen in **Figure 9**. The main sequence is the central downward trend on an HR diagram that corresponds to stars that are actively burning hydrogen. All young stars, such as members of M45, lie along the main sequence, so this is a point of evidence that the data is realistic. Some stars appear outside the main sequence, which are most likely either foreground or background stars, and not members of M45. It is also possible that these are stars that are close to the edge of the telescope's view and thus have a significant amount of streaking in the image. This would drastically affect the background subtraction aspect of the aperture photometry and may be the cause of these outliers. Another possibility is that these stars may be red giants. If this is the case, it is likely that they are foreground or background stars, as the brightest members of M45 are all B-type stars according to the Montreal Open Clusters and Associations database. As stars begin to run out of hydrogen, they cool and expand, leading to an overall rightward and upward deviation from the main sequence. In the data, these stars are brighter in red magnitudes than blue or green, which further supports this hypothesis. Many stars with lower magnitudes appear to have a large spread, which is likely caused by a low signal-to-noise ratio, causing more noise than stars with larger magnitudes. Many causes of noise are difficult to remove, especially after data has already been gathered. Despite this, it is possible to qualitatively remove some of the data points due to this telescope having a limiting magnitude of 11.2 as calculated by **Equation 4**, where M_{lim} is the limiting magnitude, and d is the telescope aperture in inches.

$$M_{lim} = 8.8 + 5log_{10}(d) \qquad \textbf{Equation 4.}^{7}$$

This limiting magnitude is the dimmest possible value that a telescope can detect with a given aperture. Taking the limiting magnitude into account by removing stars with $M_{lim} > 11.2$ eliminates a majority of this noisy region. As well as this, changing the chosen data area to exclude data that appears to be streaked significantly in the final image removes the foreground and background star region and reduces the data spread significantly, as shown by **Figure 10**.

The main sequence is visible in the resulting color-magnitude diagrams, shown in **Figure 11**. This fact shows that even with low-cost equipment, astronomical data can be retrieved with a good deal of accuracy. Entry-level astronomers can observe the stars for enjoyment while learning and understanding how astronomy is done in the actual academic field. By understanding how to go through an astronomy research project and observing good results, students can have a more positive outlook on astronomy. From this, students could be more likely to observe and produce data on other objects, such as variable stars, globular clusters, and general star-forming regions, which can be done with a low entry cost. Accessibility is key to engagement, and the easier it is to access tools to perform astronomical observations, the more likely students will be inclined to become involved with the astronomical community as a whole.

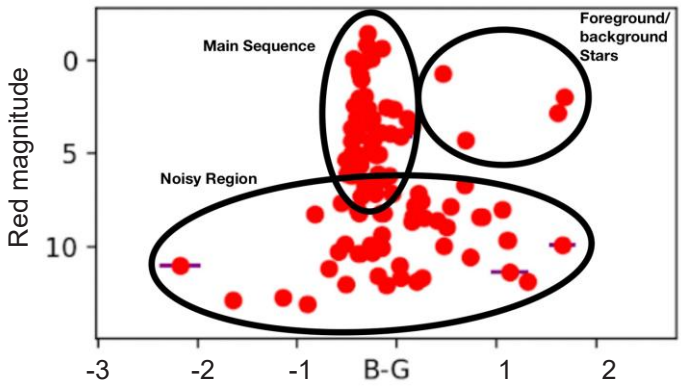


Figure 9. The apparent trend in the Red magnitude, Blue-Green color index subplot with the Main Sequence, as well as non-member/outlier stars, and noisy sources.

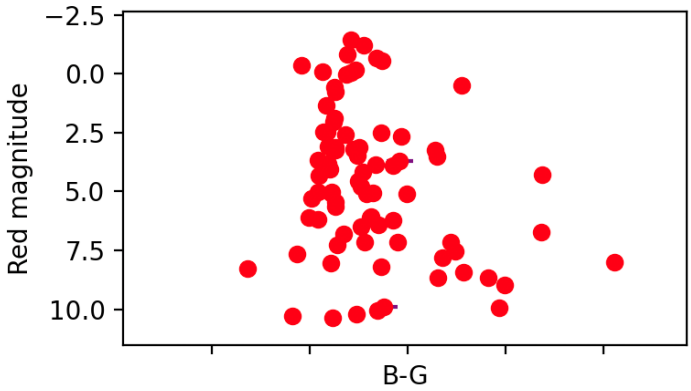


Figure 10. The same subplot as **Figure 9** with streak cropping and the magnitude limit taken into account for comparison.

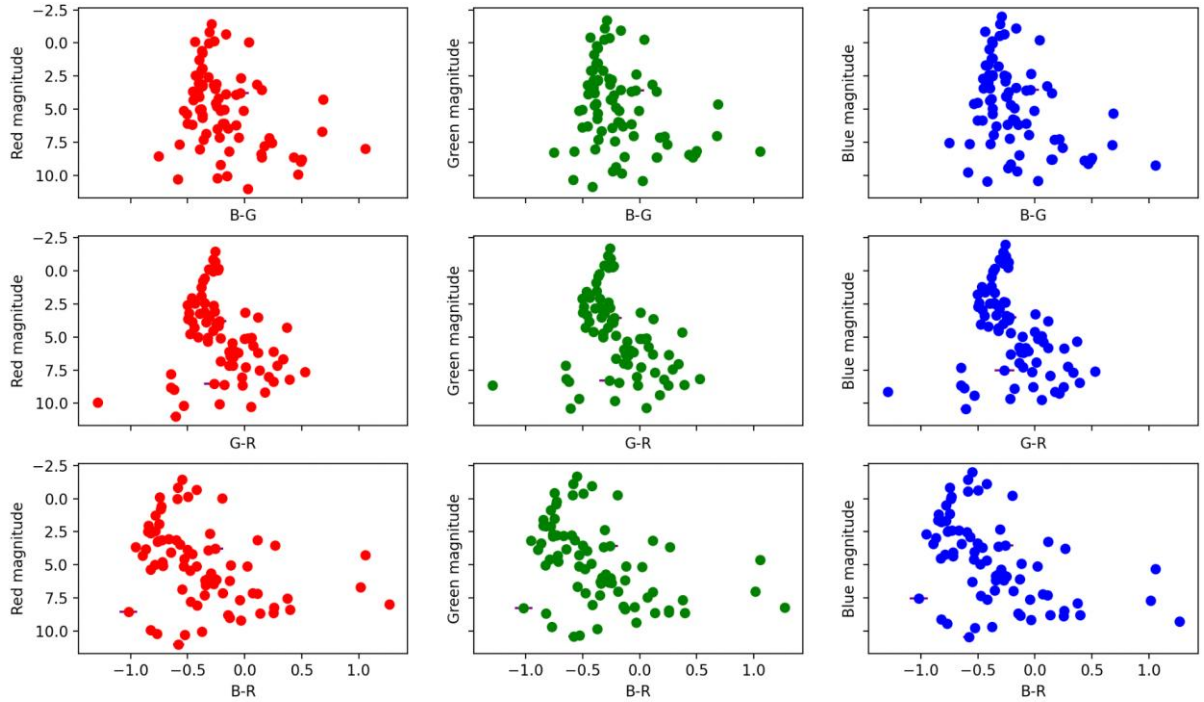


Figure 11. The resulting color-magnitude diagrams with streak cropping and magnitude limitation taken into account.

With this project, some approximations were made that could impact the accuracy of the results. For one, the observations were not done with any filters, so the color index, as previously mentioned, was blue-green rather than blue-visual, and instead of infrared, just red was used. The impact of this is likely that the magnitudes of blue and red, especially, are less extreme than they would have been had filters been used. For more accurate results, the project should be repeated with at least blue and infrared filters, which should be purchased for future iterations of this project, though one of the goals is to minimize cost, so this can be disregarded. These results also do not include any standardization to correct for extinction (absorption and scattering) caused by telluric contamination in the atmosphere known as the airmass effect. For any ground-based observation, the least amount of extinction can be obtained by observing a target along the direction of altitude of 90 degree where the airmass is defined as 1.0. Airmass value can be approximated as $Airmass = 1/cos(90° - altitude)$. The average airmass of M45 during the observations was low at 1.10, which has a definite, but relatively low impact on the results. It is also worth mentioning that the iPhone 15 Pro's camera uses a quad Bayer color filter array, meaning that a given image can be broken up into four-pixel squares where the color value is accurate. The other colors that are seen are interpolated from the surrounding pixels. The results of this project are consistent across each color, though, so the impact of this is likely small.

One thing that was quite troubling with the camera was the inability to have full control over the exposure times or ISO. The images were taken using Apple's "Night Mode" with a 5-second exposure time, but the resulting images' descriptions claimed that the maximum exposure time was 1.1 seconds. When investigating this, there was no primary documentation as to why this is. We brought the issue up to Apple photography forums, and the consensus was that the Apple camera app stops the actual data intake early for its ProRAW images to save on storage space. This is something that is troublesome and lacks a credible explanation, but it is unavoidable when using the native iPhone camera app. The iPhone's camera app also lacks direct control for the ISO, which makes it impossible to have a single ISO value for all of the images. The way to minimize the ISO was to turn the exposure down as low as it would go, but this provided an ISO of 100-400, which could theoretically be lower. Apple also uses DNG format files rather than RAW files, meaning that there was some level of processing on the images. This app was chosen because it comes predownloaded onto all iPhones, and as a result, is the most widely used camera app for iPhone users. For this reason, it is likely the app with which most people would be familiar. Despite this, it may be viable to use a different camera app that can take DNG or RAW image files without being as problematic with the exposure time or ISO. Some options are Nightcap, which is exclusive to iOS, or ProCam X, though neither is free, which would add to the overall cost of the project. If cost is not as significant an issue, one could even use an inexpensive camera that is intended for astronomical use, such as the ZWO ASI662MC USB3.0 Color Astronomy Camera. Future iterations of this research could also benefit by observing some standard astronomical object(s) and performing standard transformations on the magnitudes of the star cluster of choice to achieve higher accuracy in magnitude calculations. It would also be better in a future iteration of this project to opt not to use any digital magnification, as this diminishes the quality of the image.

Something that is also worth bringing up is the fact that flat calibration was not performed for this project. Because of this, it is possible that pixel sensitivity variation across the image is causing problems with the accuracy of the pixel brightnesses. It is also possible that the dark correction is imperfect as the exact times that the dark frames were taken was up to a few minutes apart from when some science frames were captured. To try to balance this, dark frames were taken in the middle of capturing science frames, so that the temperature when the dark frames were gathered was colder for the last science frames, but warmer than the first science frames. The conditions at the time of observation may have also been not perfectly photometric, which could impact the overall results. Even using the identical setup of this project, future students may be able to achieve a better result by implementing the full, standard digital photometric data reduction scheme used in astronomical imaging observations. This entails all the corrections already included in our project (bias, dark, stacking, etc.) plus flat correction and the standardization of the observation. The standardization step involves the observation of known brightness calibration stars, known as the photometric standard stars, to transform the measurements into the standard filter system brightness scales widely used in astronomy.

CONCLUSION

This project's broader scope can be seen in the opportunities it helps provide, especially for the astronomy education environment with a limited budget. The use of inexpensive astronomical equipment, such as the Celestron FirstScope, is an important tool for teaching the next generation of aspiring astronomers about space, astronomy research, and data science as a whole. This is supported by the apparent "main sequence" trend, as well as the low amount of error involved with the data itself. There is often a stigma with astronomy that a great deal of high-level equipment is a necessity to retrieve any reliable data. The results of this project help support the idea that this is not always the case. Students, and more generally, people interested in astronomy, can utilize low-cost equipment to perform photometry to gain a better understanding of how the brightness of stars affects our understanding of them. Astronomy, and science as a whole, should be open to anyone interested in the field and should have opportunities for those who seek them, not just those with high-grade equipment or those at the professional level. Educators should not dismiss the use of such low-cost equipment, as it holds great opportunity to be used as a tool for hands-on learning.

REFERENCES

1. Ates, Ö. & Eryilmaz, A. (2011) Effectiveness of hands-on and minds-on activities on students' achievement and attitudes towards physics. *Asia-Pacific Forum on Science Learning and Teaching, Volume 12, Issue 1, Article 6, p.1 (Jun., 2011). https://www.eduhk.hk/apfslt/v12_issue1/ates/index.htm*
2. Deslauriers, L., McCarty, L. S., Miller, K., & Kestin, G. (2019) Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(39). *https://doi.org/https://doi.org/10.1073/pnas.1821936116*
3. Kanwal, W., Shahid, S., & Jabeen, F. Effectiveness of Low Cost Material to Teach Science at Primary Level. Journal of Management Practices, Humanities and Social Sciences. *https://doi.org/10.33152/jmphss-6.2.12*
4. Harvard-Smithsonian Center for Astrophysics. (2012, April 9) *Educational materials :: Pulsating variable stars and the hertzsprung-russell diagram.* Chandra X-Ray Observatory. *https://chandra.harvard.edu/edu/formal/variable_stars/HR_student.html*

5. Center for Astrophysics Harvard & Smithsonian. (2014, Sept, 19) Retrieved February 15, 2025 from *https://www.cfa.harvard.edu/news/how-far-are-pleiades-really#:~:text=The%20Pleiades%20is%20a%20relatively,Jurassic%20dinosaurs%20passed%20into%20history*
6. Light Pollution Map. Retrieved February 15, 2025, from *https://www.lightpollutionmap.info/#zoom=15.33&lat=33.7267&lon=-83.3006&state=eyJiYXNlbWFwIjoiTGF5ZXJCaW5nUm9hZCIsIm92ZXJsYXkiOiJ3YV8yMDE1IiwibG9jIjpmYWxzZSwib3ZlcmxheW9wYWNpdHkiOjYwLCJmZWF0dXJlc29wYWNpdHkiOjg1fQ==*
7. Bowen, I. S. (1947) Notes From Observatories. *Publications of the Astronomical Society of the Pacific*, *59*, 253. *https://doi.org/10.1086/125959*

ABOUT THE STUDENT AUTHOR

Clay Reece will graduate from the University of Georgia in the Spring of 2026 with a Bachelor's degree in Astrophysics as well as a minor in Physics.

PRESS SUMMARY

Hands-on learning is crucial for students' understanding of complicated topics, such as astrophysics. Astrophysics students are often taught more theoretical concepts than practical research skills, and as a result, often do not have the opportunity to reap the benefits of hands-on education. This is a complicated problem, as it is difficult to obtain astronomical equipment due to high costs; large telescopes and good camera equipment are expensive. This is an enormous drawback for students' education, as they cannot apply the knowledge they gain in the classroom setting. This paper aims to show that producing astronomical data is possible even with low-cost, entry-level equipment, which can be utilized in classroom settings to teach students about typical astronomical research methodology through hands-on learning.

Farm Ownership, Leverage, and Government Programs Impact on Net Farm Income

Thomas Moss, Annie Kinwa-Muzinga, & Lawrence Muzinga*

Department of Applied Agricultural & Food Studies, Morningside University, Sioux City, IA

https://doi.org/10.33697/ajur.2025.160

*Student: stitchntom@gmail.com**

Mentors: kinwamuzingaa@morningside.edu, muzingan@morningside.edu

ABSTRACT

As agriculture faces increasing sustainability challenges, understanding the financial aspects of farm firms is critical for preparing future agricultural professionals. This research presents a simulation model to analyze multi-year farm profitability under different tenure structures, multiple leverage scenarios, and the presence or absence of government programs. The model considers crop prices, yields, government payments, and market fluctuations to assess the financial viability of the farm. By generating income statements, balance sheets, cash flow reports, and financial ratios, it evaluates farm stability across operations of 400, 800, and 1,200 acres. Preliminary findings suggest that government programs have a significant impact on financial resilience, risk management, and long-term profitability, with effects varying by farm size and market conditions. This study offers a valuable decision-making tool, enabling professionals to strike a balance between profitability and sustainability in an evolving agricultural landscape.

KEYWORDS

Farm Financial Analysis; Farm Profitability; Farm Tenure Structures; Government Farm Programs; Agricultural Risk Management; Financial Resilience; Farm Size & Economic Stability; Farm Firm Financial Simulation Model; Agricultural Leverage; West Central Iowa Agriculture

INTRODUCTION

The purpose of this research is to develop and validate a comprehensive farm financial simulation model. This model aims to provide agricultural professionals with insights into navigating the complexities of financial management, risk assessment, and profitability projections within the dynamic agricultural industry landscape. In agriculture, a distinct convergence of factors, including unpredictable markets, substantial capital investments, and a heavy reliance on uncontrollable variables, has created a strong demand for a comprehensive resource that empowers industry professionals to manage the complexities of their financial landscape. With such a resource, a farm financial simulation model is poised to improve capital budgeting, debt management, and profitability projections. By harnessing advanced forecasting data, comprehensive financial statements, and a deep understanding of market volatility, this simulation promises to give agricultural professionals in the West Central District of Iowa valuable insights into the future of their enterprises and the broader agricultural sector.

LITERATURE REVIEW

Bankruptcies

Several key factors emerge as critical metrics for assessing the agricultural industry's financial health and stability. These factors include net cash farm income, farm debt expense, farm real estate values, and the debt-to-asset ratio, collectively serving as indicators of profitability, asset and debt management, liquidity, solvency, and overall financial health within farm firms. The fluctuations in farm bankruptcies offer additional insight. In 2004, bankruptcies in the farm sector reached an all-time low of 0.5 per 10,000 farm firms, which marked a notable contrast to the situation in 2003, which saw an all-time high of 3.3 bankruptcies per 10,000 farm firms.[1] More recently according to a 2024 report by the American Farm Bureau Federation (2024), there were 216 Chapter 12 farm bankruptcies within 38 states, a 14% increase from the previous year, 2023, but still down 64% from 2019.[2] This uptick in Chapter 12 filings, which are specifically designed for family farmers, signals renewed financial stress in the agricultural sector, driven by factors such as high interest rates, tightening margins, and decreased commodity prices. High interest rates directly impact farmers by increasing the cost of borrowing for operating capital and long-term investments, thereby squeezing already tight profit margins. Specifically, the Midwest region has seen 71 filings, a 69% increase from 2023, while Iowa saw seven filings within the state during the same period.[2] This recent increase suggests a potential weakening of financial resilience in

specific segments of the agricultural sector, underscoring the importance of analyzing factors that contribute to farm profitability and stability.

Farm Classification

Understanding the distribution and production capacity of different farm classifications is crucial for contextualizing the financial performance and resilience that will be analyzed across 400-, 800-, and 1,200-acre farm operations, which may align with certain aspects of these broader classifications. According to the United States Department of Agriculture's (USDA) Economic Research Service, as shown in Figure 1, in 2024, small family farms accounted for 86% of all farm firms, down from 89 percent in 2021. Small family farms are typically defined as those with gross cash farm income (GCFI) of less than $350,000. Medium-sized farms comprised 6 percent, while large family and nonfamily farms constituted a 4 percent share of the total.[3] However, a mere glance at the distribution of farm firms only tells part of the story. The production values associated with these categories reveals intriguing disparities. Although small family farms dominate in numbers, they account for only 17 percent of the overall production value. Conversely, nonfamily farms, relatively scarce in number, play a significant role in the industry, accounting for 17 percent of the total production value.[3] This significant difference in production value suggests that the financial impacts of factors such as government programs and leverage may vary considerably depending on the scale of the farm operation. This disparity in production value may be attributed to factors such as economies of scale, where larger farms can achieve greater efficiency, or specialization in value-added commodities. Furthermore, different tenure structures may be more common within specific farm classifications, which could potentially influence financial outcomes. This stark contrast highlights the effectiveness of different-sized firms in utilizing available resources and emphasizes the complexity of the agricultural landscape.

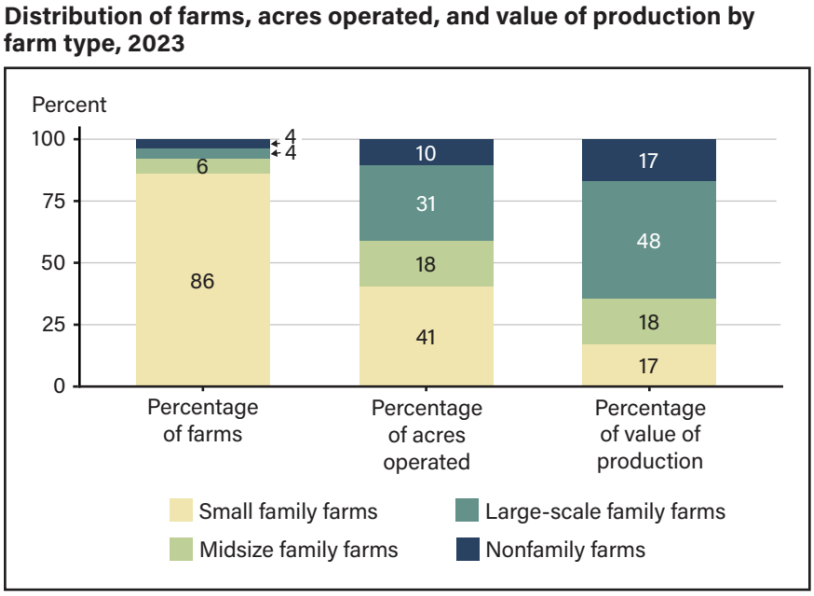


Figure 1. Farm and Ranch at a Glance 2024 Distribution of Farms. Source: United States Department of Agriculture, Economic Research Service, Farm and Ranch at a Glance 2024

Farm Ownership

Iowa's farmland has undergone significant changes over the past 40 years. The Iowa Farmland Ownership and Tenure Survey is conducted every five years, as mandated by the Iowa Code. Commentary and data are written by Tong and Zhang (2023). The survey focuses on state-level farm ownership, tenure, and transition for farmland. Changes in Iowa farmland ownership and operational structures have direct implications for farm tenure, a central aspect of this research examining farm profitability and financial stability. A few major trends have affected and will continue to affect the farming landscape in Iowa, the first being the type of ownership in which farmland is held. According to Tong and Zhang (2023) in the Iowa Farmland Ownership and Tenure Survey, the largest declines in farmland ownership are seen in the sole ownership and joint tenancy categories, which have decreased by 18% and 10%, respectively, from 1982.[4] Combined, these ownership types represented 80% of farmland in 1982 but only 52% in 2022.[4] Conversely, trusts over the same period have increased significantly, from representing only 1% of farmland in 1982 to 23% in 2022, a 2200% increase from the initial value.[4] The increasing prevalence of trusts may have implications for land transfer, potentially impacting access for new and beginning farmers, and could also influence long-term investment decisions on the land.

While the ownership structure has undergone significant changes, so have landowner demographics. The Iowa Farmland Ownership and Tenure Survey shows that Iowa farmland owners are aging at an increasing rate, particularly among those in the late stages, who are older than 74 years. In 1982, this age category owned 12% of the land, but by 2022, they now represent 37% of the total farmland, a 208% increase from the 1982 value.[4] On the other hand, early-stage owners aged 25-34 in 1982 owned 10% of the land, but by 2022, they represented only 1%, a 90% reduction from the 1982 value.[4] When combined, the age group of 65 and older has seen a 279% increase in land ownership, holding 66% of the total farmland in 2022. In contrast, the age groups 64 and younger have seen a combined decrease of 51.4% (a change from 70% in 1982 to 34% in 2022). In terms of gender distribution, males accounted for 54% of Iowa farmland owners in 2022, while females represented 46%. This proportion remains nearly unchanged from 1982, when males owned 53% and females 47%, reflecting only a 1% shift over four decades.[4] When examining the amount of time owners have owned their land, the largest share (44%) has been owned for 20 years or less, while 50 years or more represents 10% of the total land.[4] The increasing age of Iowa farmland owners raises questions about land succession, potential land availability for sale or rent, and the transfer of agricultural knowledge and practices to younger generations.

The next major trend observed in the Iowa Farmland Ownership and Tenure Survey is the operational structure of farmland, specifically whether it is owned or leased. Owned-and-operated land declined over the past 40 years, specifically going from 55% of operations to 35%, a 36.4% decrease from the initial value.[4] On the leasing side of agriculture, it is common to witness two major styles: cash rent and crop share. A cash rent agreement is one in which the farmer or tenant pays the landowner a fixed payment rate to farm the owner's land. This payment can vary depending on the type of soil and the quality of production. Under this type of agreement, the tenant assumes 100% of the risk of production and decision-making. The other major type of leasing structure is a crop share. In this type of farming agreement, the farmer and the owner share the production risk, decisions, and outcome based on their share agreement. For example, in a two-third tenant, one-third owner share agreement, the owner pays for one-third of the production costs, the tenant pays for the other two-thirds, and the production is split accordingly. The dominance of cash rent agreements, where tenants bear the full production risk, may influence their financial vulnerability and decisions regarding borrowing and investment compared to crop-share arrangements. Cash rents have become increasingly prominent over the past 40 years(1982-2022), rising from 21% to 56%, while crop shares have declined from 21% in the same year to 8% in 2022.[4] Understanding these evolving ownership and operational landscapes in Iowa is crucial for analyzing the financial performance of farm firms under different tenure structures and for evaluating the impact of government programs and leverage in this dynamic environment.

Farm size

Beyond ownership structures, the size and scale of farm operations are another crucial aspect of the agricultural sector, influencing financial performance and resilience. In Iowa, the average farm size was 355 acres in 2017, then by 2022, this figure had dropped to 345 acres, marking only the second time that average farmland had decreased over the last 82 years year over year (1940-2022), with the only other drop being reported in 2002.[5] It is important to note that the 2002 drop is credited to the USDA changing reporting requirements, which included smaller farms by decreasing the production minimum. The farm sizes modeled in this research (400, 800, and 1,200 acres) generally represent operations larger than the current average in Iowa, allowing for an analysis of financial dynamics in more substantial farming enterprises. The slight decrease in average farm size in Iowa, contrasting with the national trend, may be attributed to factors such as higher land values making expansion more difficult, an increase in smaller specialty farms, or land being divided for generational transfer. This contrasted with the national average for the United States, which in 2017 was 441 acres.[5] However, by 2022, it had grown by 22 acres to reach 463 acres, and the number of farms had decreased by 141,733.[5] The scale of a farm operation can significantly influence the choice of tenure structure, the amount of leverage a farm can manage, and the applicability or benefits derived from various government programs.

The rising value of farm assets, mainly farmland, is a notable factor. For the four years from 2018 to 2021, the value of one acre of farmland in Iowa appreciated by 57.09%, surging from $7,264 to $11,411.[6] Notably, in 2023, land values in Iowa continued to appreciate , reaching $11,835 before falling in 2024 to $11,467, a 3.1% year-over-year decline, marking the first drop in value since 2018.[6] The rapid appreciation and subsequent slight decline in Iowa farmland values create a dynamic financial environment, impacting farmers' asset wealth, borrowing capacity, and the overall risk associated with agricultural investments.

Government programs

Government programs represent another critical layer that influences the financial stability and profitability of agricultural operations. Government payments have long played a role in supporting financial stability within U.S. agriculture, particularly for farmers in Iowa. However, these payments have never been static in their structure, calculation methods, and policy objectives evolving considerably over time.

A recent major shift occurred with the passage of the Agriculture Act of 2014, which eliminated the direct payment system that had been in place since 1996, as stated by the United States Senate Committee on Agriculture.[7] During that period, according to Purdue University Center for Commercial Agriculture, direct payments were designed to help farmers transition to a market-oriented system by providing predictable income support as agriculture moved away from government intervention and toward supply-and-demand pricing. When this direct payment system was discontinued, it was replaced by a variety of new farm subsidy programs, each with different eligibility and calculation criteria.[8] Over the years, these programs have continued to change, resulting in significant variability in payment levels.

Data from the USDA Economic Research Service (ERS) Farm Income and Wealth Statistics (2025), expressed in 2025 real dollars, illustrates historical fluctuations in government payments. The graph below focuses exclusively on payments related to crop production, ad hoc programs, and disaster assistance, as these categories most accurately reflect conditions for West-Central Iowa farmers modeled in this research. Using 2025 real dollars ensures that all values are adjusted for inflation, allowing for consistent comparison of payment levels across different years. Iowa crop producers received their highest payments in 2020, totaling approximately $3.84 million, driven mainly by supplemental and ad hoc disaster assistance. Historical trends reveal dramatic swings as seen between 1977 and 1978, payments surged by 10,705.61%, rising from $1,728 to $186,721. Conversely, in 1995, payments fell by 96.48% compared to the previous year, dropping from $578,339 in 1994 to just $20,335. The graph below illustrates total government program payments and crop-specific payments in Iowa from 1949 to 2024, adjusted to 2025 real dollars.[9]

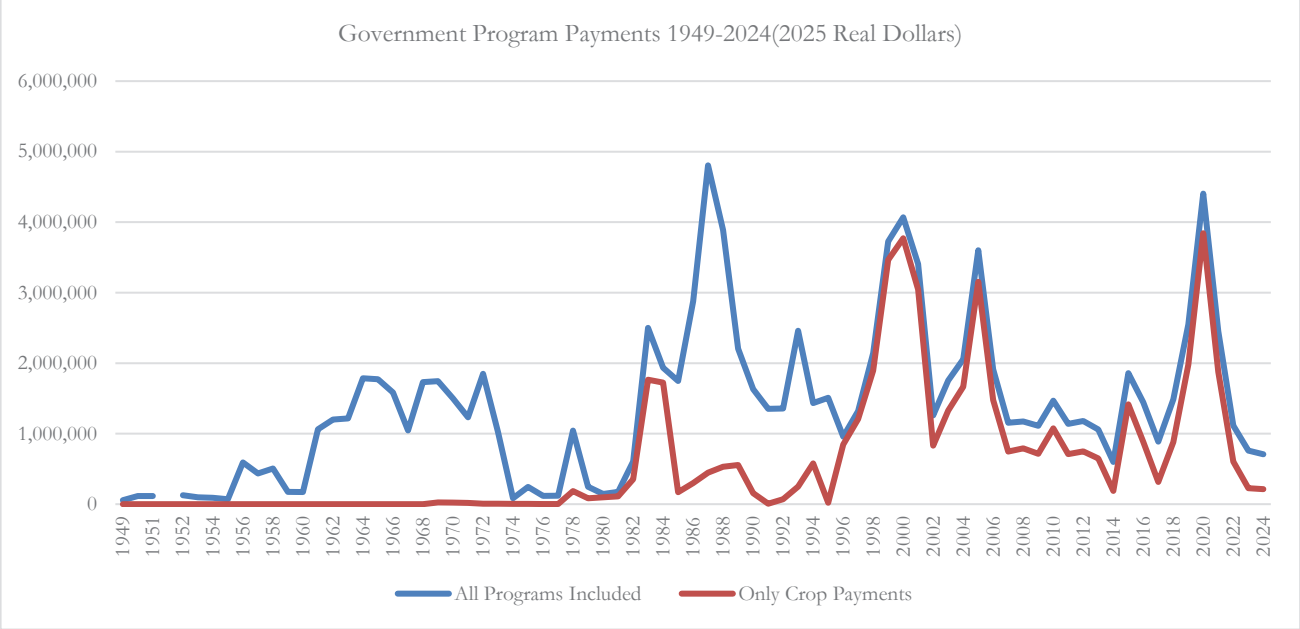


Figure 2. Iowa Government Program Payments 1949-2024 Source: Authors.

When looking at the USDA's net farm income forecast for the 2025 growing year, it could be assumed that the agricultural industry is healthy and making a surge from the 2024 growing year, rising by $180.1 billion or 29.5%, as seen in Figure 2.[10] This forecasted increase highlights the impact of government programs despite declining crop receipts on net farm income. Crop receipts, specifically corn, are projected to decline by 4.3%, and soybeans are expected to fall by 6.6% from 2024 to 2025.[10] With crop receipts expected to fall in 2025, the USDA's projection is out of context until government programs are factored in. In 2025, government programs are estimated to account for $42 billion or 23.6% of the total net farm income.[10] When comparing the 2025 estimates, in 2024, government programs accounted for 6.5%; in 2023, they were 8.2%; in 2022, they were 8.8%; and in 2021, they were 17.8%.

The major driver of this increase in 2025 is the American Relief Act of 2025, an ad hoc program passed in December of 2024 in response to ongoing economic challenges in the agricultural sector, providing significant disaster and financial assistance to producers. This program includes $31 billion to be distributed between two central pillars. The first $21 billion is for disaster relief, and the second $10 billion is for economic relief to producers. Given the significant and variable contribution of government programs to net farm income, as highlighted by the 2025 forecast, this research incorporates the presence or absence of such programs as a key variable in the simulation model to assess their impact on farm profitability and financial resilience.[10]

Along with this bill, Price Loss Coverage (PLC) and Agricultural Risk Coverage (ARC) are expected to triple from 2024 to 2025 due to the decline in commodity prices. The simulation model will specifically analyze the impact of programs such as Price Loss Coverage (PLC) and Agricultural Risk Coverage (ARC), which are projected to see increased payouts in 2025. It is plausible that the benefits and accessibility of these government programs may vary depending on a farm's tenure structure (owned vs. leased) or its level of financial leverage, aspects that will be explored in this research. The American Farm Bureau (2024) notes that while both the ARC and PLC payments provide a safety net for financial stability, the programs lack mitigation tactics for trade risk.[2]

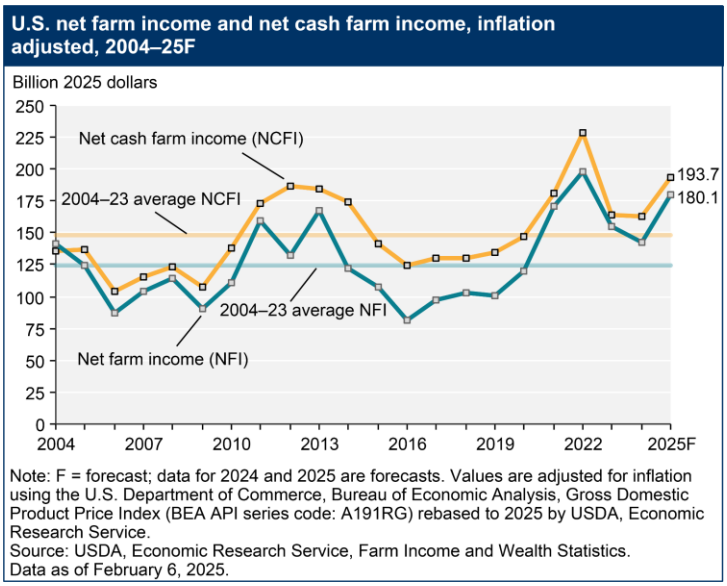


Figure 3. USDA Net Farm Income Projection 2025

Farm Debt

Another critical aspect of the agricultural financial landscape is the level and management of farm debt, which can significantly influence a farm's vulnerability to economic downturns. Debt trends can be first identified by tracking banking information and trends in agricultural lending. When analyzing the Federal Reserve Bank of Kansas City's (2025) 2024 fourth-quarter commercial bank report, it is seen that in 2024, farm debt outstanding rose at agricultural banks by 10% on non-real estate loans and 4% on real estate loans, contrary to non-agricultural banks that saw no increase from the previous year in real estate loans but saw a 2% increase in non-real estate loans.[11] These trends in rising farm debt at agricultural banks highlight the critical importance of analyzing the impact of different leverage scenarios on farm financial stability, a key component of this research.

Agricultural banks are well above historical 20-year averages in both real estate and non-real estate loans. Delinquency rates increased from historic lows in both real estate and non-real estate loans, climbing above 1% for the first time since 2020. Specifically, the duration of delinquency over 90 days more than tripled year-over-year.[12] The increase in delinquency rates and the reduced likelihood of some farmers qualifying for operating credit suggest potential financial headwinds in the agricultural sector, possibly influenced by factors such as declining commodity prices and increasing input costs.

Complementing these trends is the Federal Reserve Bank of Chicago's February 2025 AgLetter, which found that 44% of survey respondents reported an increase in non-real estate farm loans from the previous year, with only 15% reporting a decrease. Along with this, 1.7% of farm customers who had operating credit last year were not likely to qualify for new operating credit in 2026. This may disproportionately affect new and beginning farmers who lack a long credit history or substantial asset base. Notably, 40% of banks reported tightening credit standards for farm loans, and 19% noted an increase in the amount of loan collateral.[13] These financial trends reported by the Federal Reserve Banks are particularly relevant to the Iowa agricultural sector, which is the focus of the simulation model. Access to and utilization of farm debt may also vary depending on the farm's tenure structure, as landowners may have different collateral options compared to tenant farmers.

Importance

The forthcoming financial simulation model from this research is poised to have a significant impact on the agricultural community it serves. By comprehensively understanding the industry's historical context and prevailing trends, this model is engineered to empower and elevate the financial decision-making processes of agricultural professionals.

The model relies heavily on both historical and current forecasting data. The paramount consideration in data collection is prioritizing localized information, as it is most relevant to the end-users within the community. Furthermore, the model draws its data from a roster of trusted and respected sources, including but not limited to the USDA (United States Department of Agriculture), NASS (National Agriculture Statistic Services), Iowa State University, University of Missouri, and other reputable entities.

METHODS AND PROCEDURES

Scenarios

In this research, the variables of acres, leverage, tenure, and government programs were analyzed to determine the maximized operation based on net farm income per acre. The farm size levels (400, 800, and 1,200 acres) were selected to represent a range of small to large farm sizes common in West Central Iowa, while the leverage levels (10%, 25%, 50%, and 75%) span from low to high debt-to-asset ratios observed in agricultural operations. Government programs were either included in the simulation or excluded from the simulation to show the effects of these overall payments. When government programs were factored in, the model included payments from programs such as Price Loss Coverage (PLC), Agricultural Risk Coverage (ARC), and ad hoc disaster relief payments as outlined in the 2025 forecasts.

The primary metric for determining the maximized operation was net farm income per acre. However, the simulation also generated income statements, balance sheets, cash flow reports, and key financial ratios for a comprehensive assessment of economic viability and stability under each scenario.

Simulation Scenarios			
Scenario #	Ownership(Tenure) %	Cash Rent %	Crop Share %
1	100	0	0
2	75	100	0
3	75	75	25
4	75	50	50
5	75	25	75
6	75	0	100
7	50	100	0
8	50	75	25
9	50	50	50
10	50	25	75
11	50	0	100
12	25	100	0
13	25	75	25
14	25	50	50
15	25	25	75
16	25	0	100
17	0	100	0
18	0	75	25
19	0	50	50
20	0	25	75
21	0	0	100

Table 1. Simulation tenure scenarios. Source: Authors.

Simulation

This type of analysis allows representative farms in the West Central District of Iowa to: (1) assess the riskiness of their decisions, (2) plan appropriately for each possible outcome, (3) gain a better understanding of the farming operation process, (4) acknowledge the uncertainty of the results, and (5) present a more credible and justifiable set of decisions.[14]

Like most simulation models, this one is solved over a range of deterministic input values. This study assumes that crop prices and yields are deterministic over a 5-year period. Crop prices and yields were modeled using historical distributions fitted to recent market data obtained from USDA-NASS for Iowa (2014-2023).[15] The parameters of these distributions (mean and standard deviation) were derived from this data. For the first year, family consumption was modeled using a normal distribution based on the average family living expenses for farm households in Iowa as reported by the USDA ERS (2023).[14] Key deterministic variables in the model include interest rates based on the defined leverage levels, government program payment rates as per the 2023 Farm Bill, and projections for ad hoc payments and baseline operating costs derived from Iowa State University Extension data (2024).[6] For each of the 504 distinct scenarios (3 acre sizes, 4 leverage levels, 2 government program conditions, 21 tenure structures), the simulation was executed.

Farm Simulation Assumptions

The research objective of developing a farm profitability forecasting model was successfully achieved by establishing key assumptions to delineate the profile of the farm firm and its operating environment. These assumptions create the foundation for the simulation, enabling a comprehensive understanding of the farm firm's dynamics before external factors are introduced.

Financially, the farm firm possesses $64,843 in cash reserves, a figure based on the average liquid asset holdings of farms in the Midwest with similar acreage.[16] A minimum cash reserve requirement of $30,000 was set to ensure the farm firm maintains sufficient liquidity to cover short-term obligations and unexpected expenses within the simulation. Additionally, the firm allocated $60,000 to marketable securities, distributed as 12.5% in the S&P 500, 12.5% in bonds, and 75% in Certificates of Deposit (CDs), representing a diversified investment strategy common among farm operations to manage retained earnings. Another $60,000 was allocated to longer-term, less liquid investments held outside of daily trading, with a distribution of 50% in the S&P 500 index funds, 25% in bond funds, and 25% in Certificates of Deposit. These will be referred to as "non-marketable securities" within the model. The initial levels of cash reserves and marketable securities can affect a farm's capacity and willingness to utilize debt financing under the different leverage scenarios.

The simulation also incorporates detailed crop input costs on a per-acre basis. These costs include fertilizer, herbicides, insecticides, seed, drying and storage, machinery hire and repairs, labor, and other crop-related expenses. On average, corn production costs in Year 0 total $338.27 per acre, while soybean production costs average $206.63 per acre. In addition, the analysis accounts for property taxes, with an assumed average of $27.81 per cropland acre also in Year 0.

Furthermore, the simulation incorporated growth rates to project changes in costs and asset values over the 5-year forecasting period. Notably, crop costs were projected to increase at an annual rate of 2.17%, based on the average of USDA projections for input costs.[17] The Consumer Price Index (CPI) was expected to grow by 4.5% annually, reflecting the average of the USDA Chief Economist's 2032 (2023) economic projections. Real estate values were forecasted to appreciate at a rate of 3.75% annually, based on the historical trends in Iowa farmland values reported by the Iowa State University Land Value Survey.[6] Marketable securities were anticipated to grow at 6.17% annually, representing the long-term historical average return of a balanced portfolio of S&P 500 and bonds.[18] Non-marketable securities were projected to grow at a 9.3% annual rate, reflecting a potentially more aggressive long-term investment strategy. The underlying trend for crop costs is assumed to follow this deterministic growth rate. In scenarios involving crop share agreements, a 67% split was assumed for the tenant and a 33% split for the landowner.

These assumptions collectively formed the basis of the farm profitability forecasting model, facilitating the generation of realistic and insightful financial projections over the specified timeframe. Unless otherwise stated, all financial results are expressed from the viewpoint of the tenant farmer.

RESULTS

Crop Share vs Cash Rent

One of the major findings in this research is the significant difference in net farm income returned to the farm tenant observed between 100% crop share and 100% cash rent tenure structures across varying leverage levels and farm sizes. By comparing all scenarios with 100% crop share (scenarios 6, 11, 16, 21) to scenarios with 100% cash rent (scenarios 2, 7, 12, 17), it was found that, on average, crop share resulted in a higher net farm income per acre at lower leverage levels. Specifically, at a 10% leverage, the average net farm income per acre was $89.81 higher under 100% crop share compared to 100% cash rent. This difference increased to $123.18 per acre at 25% leverage and $146.27 per acre at 50% leverage. At the highest leverage level of 75%, the average difference in net farm income per acre between crop share and cash rent was $155.16, still favoring crop share on average.

These average differences were calculated by taking the mean net farm income per acre across all 100% crop share scenarios and comparing it to the mean across all 100% cash rent scenarios within each specified leverage level, with government programs factored into the simulation.

Interestingly, the farm size at which the largest change in net farm income occurred varied depending on the leverage. At the 10%, 25%, and 50% leverage scenarios, the largest positive impact of crop share over cash rent was observed on the 400-acre farm, with an average difference of $96.48 per acre. However, at the 75% leverage level, the most considerable difference shifted to the 800-acre farm. Conversely, the smallest difference in net farm income per acre between the two tenure structures across all leverage scenarios was consistently found in the 1,200-acre farm.

This suggests that at lower leverage, the returns under a crop share agreement outweigh the shared risk and lower operating costs associated with crop share, particularly on smaller farms. However, as leverage increases, the financial implications of fixed cash rent versus crop share have a more pronounced impact, potentially interacting differently with farm size and debt obligations. These differences are further illustrated in Figure 3, which provides a detailed breakdown of net farm income per acre by leverage level, which is averaged across all farm sizes.

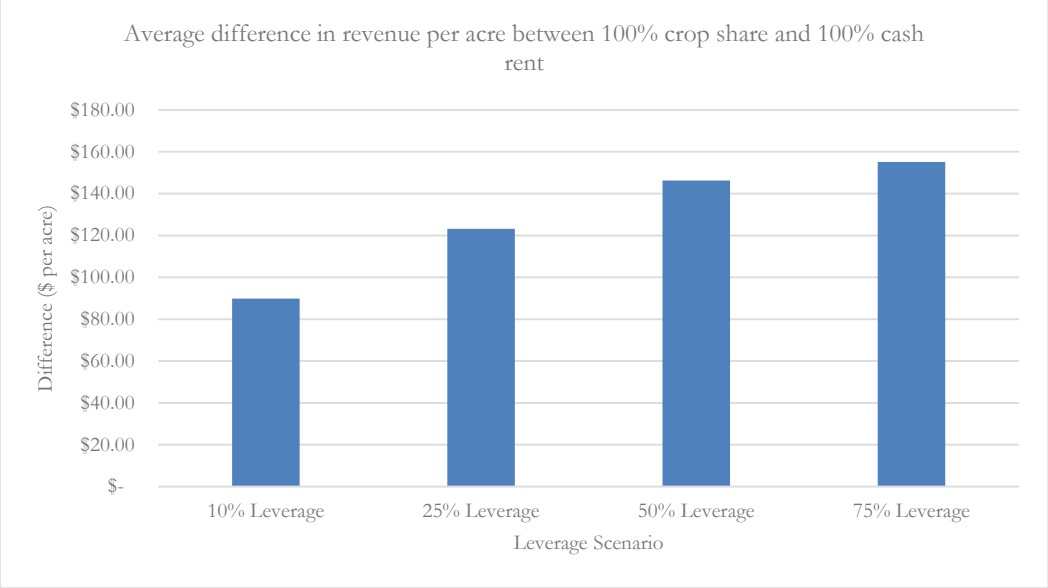


Figure 4. Revenue Difference Between 100% Crop Share and 100% Cash Rent. Source: Authors.

When comparing the 50% crop share and 50% cash rent scenarios (Scenarios 4, 9, 14, and 19), profitability varies across leverage and ownership levels. At 10% leverage, the highest net income per acre occurs on a 400-acre farm with 75% ownership, while the lowest is on a 1,200-acre farm with 0% ownership. At 25% leverage, the most profitable scenario shifts to a 400-acre farm with 0% ownership, whereas the least profitable is a 1,200-acre farm with 75% ownership. At 50% leverage, the highest income remains with a 400-acre farm at 0% ownership, and the lowest again is a 1,200-acre farm at 75% ownership. Finally, at 75% leverage, the pattern persists where the highest income is achieved by a 400-acre farm with 0% ownership, and the lowest by a 1,200-acre farm owning 75% of its land.

This trend reflects the interaction between leverage and ownership costs. At higher leverage levels, farms with 0% ownership perform better by avoiding interest expenses and other fixed costs associated with land ownership. Conversely, at lower leverage levels, farms with higher ownership tend to show greater returns due to reduced ongoing rental costs.

Government programs

The simulation results clearly demonstrate that government programs significantly increase net farm income across all farm sizes and leverage levels. It was found that the difference between having government programs on a farm, with 10% leverage and farming 400 acres is $155.17 higher with programs. The difference is $145.53 higher for 800 acres and $141.19 higher for 1,200 acres. This positive impact of government programs grows as leverage increases. A 400-acre farm sees a $199.53 increase, an 800-acre farm a $199.72 increase, and a 1,200-acre farm a $199.60 increase.

On average across all leverage levels and farm sizes, 27 scenarios witnessed a negative net farm income per acre with government programs in place. In contrast, the average increased to 56 scenarios without government programs, particularly at the 50% and 75% leverage rates, especially with high land ownership. This suggests that at higher debt burdens and when a larger proportion of land is owned (resulting in higher fixed costs such as mortgage payments and property taxes), the absence of government

support significantly increases the vulnerability to negative profitability. Interestingly, the positive impact of government programs on net farm income also varied across different tenure structures, with leased land sometimes showing a slightly higher dependence on programs at high leverage, as the operator bears the full production risk under cash rent agreements. The significant impact of government programs is visually represented in Figure 4, which compares the distribution of net farm income per acre with and without program payments across different leverage levels and farm sizes.

Figure 5. Impact of Government Programs on Net Farm Income Across Different Leverage Levels. Source: Authors.

Scenario Analysis

When government programs were included, notable effects on net farm income were observed. The 800-acre farm was able to maximize net farm income with 100% land ownership at 10% leverage, leading to a $278.42 per acre income. The lowest net farm income conversely was on a 100% land ownership farm at a 75% leverage level with government programs, leading to a -$566.53 per acre loss as seen below in Table 2. A 400-acre operation follows the same key scenarios, where maximizing net farm income per acre with 100% land ownership at 10% leverage led to a $343.53 per acre income. This is a difference of $904.05 from the lowest net farm income of -$560.52 as seen in Table 3. At the 1,200-acre operation, the highest net income was $253.05 per acre, and the lowest was -$568.53 per acre.

800 Acre- Revenue Per Acre Including Government Programs				
	10%	25%	50%	75%
1	$ 278.42	$ 149.06	$ (206.88)	$ (566.53)
2	$ 243.00	$ 129.02	$ (152.70)	$ (436.83)
3	$ 250.37	$ 143.36	$ (137.33)	$ (421.11)
4	$ 257.88	$ 157.54	$ (121.96)	$ (405.39)
5	$ 265.33	$ 171.55	$ (106.58)	$ (389.67)
6	$ 272.46	$ 185.48	$ (91.21)	$ (373.96)
7	$ 200.94	$ 108.77	$ (98.52)	$ (307.13)
8	$ 218.87	$ 137.65	$ (67.78)	$ (275.69)
9	$ 235.80	$ 165.90	$ (37.03)	$ (244.26)
10	$ 251.23	$ 193.76	$ (6.29)	$ (212.82)
11	$ 266.14	$ 218.51	$ 24.46	$ (181.38)
12	$ 153.50	$ 88.19	$ (44.35)	$ (177.43)
13	$ 182.71	$ 131.86	$ 1.77	$ (130.27)
14	$ 210.72	$ 173.45	$ 47.89	$ (83.12)
15	$ 236.61	$ 206.64	$ 93.28	$ (35.96)
16	$ 259.50	$ 232.45	$ 137.74	$ 11.19
17	$ 99.36	$ 67.39	$ 9.83	$ (47.73)
18	$ 144.48	$ 125.73	$ 71.28	$ 15.14
19	$ 183.52	$ 171.54	$ 131.06	$ 77.84
20	$ 220.61	$ 210.14	$ 187.52	$ 138.67
21	$ 252.76	$ 243.05	$ 227.34	$ 197.23

Table 2. 800 Acre- Revenue Per Acre Including Government Programs. Source: Authors.

400 Acre- Revenue Per Acre Including Government Programs				
	10%	25%	50%	75%
1	$ 343.53	$ 187.49	$ (182.91)	$ (560.52)
2	$ 305.29	$ 167.75	$ (128.74)	$ (430.82)
3	$ 314.28	$ 181.84	$ (113.36)	$ (415.10)
4	$ 321.98	$ 195.78	$ (97.99)	$ (399.38)
5	$ 330.17	$ 209.70	$ (82.61)	$ (383.66)
6	$ 336.99	$ 223.63	$ (67.25)	$ (367.95)
7	$ 258.64	$ 147.87	$ (74.56)	$ (301.12)
8	$ 278.09	$ 176.20	$ (43.81)	$ (269.68)
9	$ 297.15	$ 204.05	$ (13.07)	$ (238.25)
10	$ 315.13	$ 231.91	$ 17.35	$ (206.81)
11	$ 330.92	$ 259.77	$ 47.65	$ (175.37)
12	$ 208.54	$ 127.73	$ (20.38)	$ (171.42)
13	$ 240.07	$ 170.54	$ 25.41	$ (124.26)
14	$ 269.23	$ 212.33	$ 70.68	$ (77.11)
15	$ 297.99	$ 253.88	$ 115.39	$ (29.95)
16	$ 324.21	$ 289.61	$ 159.32	$ 16.92
17	$ 150.25	$ 107.45	$ 33.46	$ (41.72)
18	$ 200.92	$ 164.83	$ 93.69	$ 20.89
19	$ 240.92	$ 220.12	$ 152.74	$ 82.65
20	$ 279.72	$ 264.38	$ 210.03	$ 143.26
21	$ 316.52	$ 302.77	$ 267.03	$ 202.02

Table 3. 400 Acre- Revenue Per Acre Including Government Programs. Source: Authors.

1,200 Acre- Revenue Per Acre Including Government Programs				
	10%	25%	50%	75%
1	$ 253.05	$ 136.19	$ (214.87)	$ (568.53)
2	$ 220.17	$ 115.98	$ (160.69)	$ (438.83)
3	$ 227.30	$ 130.40	$ (145.32)	$ (423.11)
4	$ 234.13	$ 144.68	$ (129.94)	$ (407.39)
5	$ 240.70	$ 158.71	$ (114.57)	$ (391.68)
6	$ 247.24	$ 171.78	$ (99.20)	$ (375.96)
7	$ 180.77	$ 95.49	$ (106.51)	$ (309.13)
8	$ 198.48	$ 124.61	$ (75.77)	$ (277.69)
9	$ 213.90	$ 153.17	$ (45.02)	$ (246.26)
10	$ 228.11	$ 179.60	$ (14.28)	$ (214.82)
11	$ 241.47	$ 200.61	$ 16.47	$ (183.39)
12	$ 133.98	$ 74.76	$ (52.34)	$ (179.43)
13	$ 163.43	$ 118.82	$ (6.21)	$ (132.28)
14	$ 190.70	$ 158.49	$ 39.90	$ (85.12)
15	$ 214.59	$ 187.68	$ 85.84	$ (37.96)
16	$ 235.63	$ 212.01	$ 130.50	$ 9.19
17	$ 81.30	$ 53.42	$ 1.84	$ (49.73)
18	$ 124.83	$ 110.90	$ 63.33	$ 13.14
19	$ 164.28	$ 153.58	$ 123.78	$ 76.01
20	$ 200.23	$ 190.85	$ 173.81	$ 137.14
21	$ 229.61	$ 222.30	$ 208.71	$ 192.06

Table 4. 1,200 Acre- Revenue Per Acre Including Government Programs. Source: Authors.

This analysis leads to the conclusion that at a 10% leverage rate, net farm income is maximized when land ownership is 100% across all acreage scenarios. At a 25% leverage rate income maximization varies based on acres, where a 1,200-acre farm's income will be maximized at 50% land ownership, an 800-acre farm will maximize income at 25% land ownership, and finally a 400-acre farm will have 0% land ownership to maximize income. Finally, at a 50% and 75% leverage rate across all acre farms, 0% land ownership will maximize net income.

DISCUSSION

This research analyzed the impact of farm size, leverage, tenure, and government programs on farm profitability, revealing key insights. Firstly, despite representing only 8% of farmland, 100% crop share arrangements in our simulation yielded a significant 31%-58% higher net farm income per acre than 100% cash rent across all leverage and acre scenarios. This suggests that the benefits of shared risk and flexible costs in crop share may be undervalued, warranting further investigation into the drivers behind the dominance of cash rent and the characteristics of land under crop share. In addition to understanding the factors driving shifts in leasing structures from both tenant and landowner perspectives, it is essential to model these arrangements using both deterministic and stochastic approaches. This will allow for a comprehensive evaluation of the financial impact of various lease types and differing levels of shared risk.

Secondly, the simulation underscores the crucial role of government programs in upholding farm income, particularly under high leverage and land ownership where fixed costs are substantial. This aligns with projections showing increased reliance on government payments. However, the long-term sustainability and market implications of this dependence require further consideration. Future research could focus on optimizing program design to achieve a balance between support and market efficiency.

Finally, an inverse relationship emerged between leverage and optimal land ownership. Higher leverage often favors lower ownership levels, likely due to the increased burden of fixed ownership costs, including interest and taxes, compared to the more flexible cost structures of leasing, especially crop share. This highlights the need to carefully align tenure decisions with debt

levels. Future research could build on these findings by incorporating the data into a stochastic modeling framework to evaluate the degree of correlation between simulated outcomes and actual historical results. This approach would provide insight into the reliability and predictive accuracy of the model under varying risk and uncertainty conditions.

CONCLUSIONS

The farm financial simulation model developed in this study was used to evaluate how farm size, leverage, tenure structure, and government programs influence profitability in the West-Central district of Iowa, yielding several key insights. First, crop share arrangements consistently outperformed the more common cash rent model. This finding suggests the need for future research to explore the motivations behind leasing preferences and to develop models that identify the most equitable distribution of risk and return between landowners and tenants. Second, government programs were shown to have a critical impact on farm income, particularly for highly leveraged operations. Finally, the analysis revealed an inverse relationship between leverage and optimal land ownership, indicating that higher debt levels favor lower ownership percentages.

While these findings provide a valuable framework for decision-making, they also raise important questions for future research. Further studies could incorporate stochastic modeling to assess the reliability of these results under uncertainty and compare simulated outcomes with historical performance. Additionally, exploring the long-term sustainability of government support and its interaction with market efficiency, as well as analyzing more nuanced tenure structures and risk-sharing arrangements, would deepen understanding of farm financial resilience. By advancing these areas, future research can refine strategies that balance profitability, risk, and sustainability in an evolving agricultural economy.

These findings provide a strong foundation for future research, but there are several areas that warrant deeper exploration. Future studies could incorporate stochastic modeling to capture variability and uncertainty, producing more robust and predictive results. Additionally, recreating this model for historical periods would allow for validation and verification of its accuracy over time. Beyond model refinement, examining the long-term effects of government programs and payments on market dynamics could inform policy decisions aimed at creating a sustainable and competitive agricultural marketplace. By advancing these research directions and maintaining a focus on risk management, profitability, and economic sustainability, farm firms in West-Central Iowa will be better equipped with tools to achieve both financial resilience and environmental stewardship.

REFERENCES

1. Key, N., Law, J., & Whitt, C. (2021, November) Chapter 12 bankruptcy rates have increased in most agricultural states. USDA Economic Research Service. *https://www.ers.usda.gov/amber-waves/2021/november/chapter-12-bankruptcy-rates-have-increased-in-most-agricultural-states/*
2. American Farm Bureau Federation. (2024, March 7) 2024 farm bankruptcies highlight worsening farm credit. Farm Bureau Market Intel. *https://www.fb.org/market-intel/2024-farm-bankruptcies-highlight-worsening-farm-credit*
3. United States Department of Agriculture, Economic Research Service. (2024, December) America's farms and ranches at a glance: 2024 edition (EIB-283). United States. Department of Agriculture. *https://ers.usda.gov/sites/default/files/_laserfiche/publications/110560/EIB-283.pdf*
4. Tong, J., & Zhang, W. (2023, August) Iowa farmland ownership and tenure survey 1982–2022: A forty-year perspective (Working Paper 23-WP 651). Center for Agricultural and Rural Development, Iowa State University. *https://farmland.card.iastate.edu/files/inline-files/farmland-ownership-tenure-2022_0.pdf*
5. United States Department of Agriculture, National Agricultural Statistics Service. (2024) 2022 Census of Agriculture: Iowa State and County Data (Volume 1, Geographic Area Series, Part 15) (AC-22-A-15). United States Government Publishing Office. *https://www.nass.usda.gov/Publications/AgCensus/2022/Full_Report/Volume_1,_Chapter_1_State_Level/Iowa/st19_1_001_001.pdf*
6. Iowa State University Extension and Outreach. (2024) 2024 Farmland Value Survey (AgDM File C2-70). Iowa State University. *https://www.extension.iastate.edu/agdm/wholefarm/html/c2-70.html*
7. United States Senate Committee on Agriculture, Nutrition, and Forestry. (2014) Farm bill ends direct payment subsidies. United States Senate. *https://www.agriculture.senate.gov/newsroom/press/release/farm-bill-ends-direct-payment-subsidies*
8. Purdue University Center for Commercial Agriculture. (2014) The end of the direct payment era in U.S. farm policy. Purdue University. *https://ag.purdue.edu/commercialag/home/resource/2014/01/the-end-of-the-direct-payment-era-in-us-farm-policy*
9. United States Department of Agriculture, Economic Research Service. (2025) Farm income and wealth statistics. United States Department of Agriculture. *https://www.ers.usda.gov/data-products/farm-income-and-wealth-statistics/data-files-us-and-state-level-farm-income-and-wealth-statistics*
10. United States Department of Agriculture, Economic Research Service. (2024, May 22) Farm sector income forecast. United States Department of Agriculture. *https://www.ers.usda.gov/topics/farm-economy/farm-sector-income-finances/farm-sector-income-forecast/*
11. Federal Reserve Bank of Kansas City. (2025) Ag Finance Databook. Federal Reserve Bank of Kansas City. *https://www.kansascityfed.org/agriculture/ag-finance-databook/*

12. Scott, F., & Kreitman, T. (2025, March 27) Farm debt at ag banks climbed in 2024. Federal Reserve Bank of Kansas City. *https://www.kansascityfed.org/agriculture/agfinance-updates/farm-debt-at-ag-banks-climbed-in-2024/*

13. Federal Reserve Bank of Chicago. (2025, February 13) AgLetter: February 2025. *https://www.chicagofed.org/publications/agletter/2025-2029/february-2025*

14. Muzinga, N. P. (2005) Modeling Farm-level impacts of Federal Income Tax Reforms: A stochastic simulation approach. IDEALS. *https://hdl.handle.net/2142/82974*

15. USDA-NASS for Iowa (2014-2023) (n.d.) In Iowa agricultural statistics. Retrieved August 22, 2025, from *https://www.nass.usda.gov/Statistics_by_State/Iowa/index.php*

16. United States Department of Agriculture, Economic Research Service. (2024) Agricultural baseline projections. United States Department of Agriculture. *https://www.ers.usda.gov/topics/farm-economy/agricultural-baseline-projections/*

17. United States Department of Agriculture, Office of the Chief Economist. (2023) USDA agricultural projections to 2032 (OCE-2023-01). United States Department of Agriculture. *https://www.usda.gov/sites/default/files/documents/USDA-Agricultural-Projections-to-2032.pdf*

18. Morningstar. (2024, January 18) Experts forecast stock and bond returns: 2024 edition. Morningstar. *https://www.morningstar.com/portfolios/experts-forecast-stock-bond-returns-2024-edition*

ABOUT THE AUTHOR

Thomas (Tom) Moss is a professional farm and ranch manager and licensed real estate agent in Nebraska. He graduated in 2025 with honors from the Department of Applied Agricultural & Food Studies and the School of Business at Morningside University in Sioux City, Iowa, earning a Bachelor of Science in Applied Agricultural & Food Studies and a Bachelor of Arts in Business Administration. His undergraduate research focused on agricultural economics and finance, examining how farm tenure, leverage, and government programs influence farm profitability. For his honors thesis, he developed a simulation-based stochastic forecasting model for farm profitability in West Central Iowa. In his professional work, he applies financial modeling, income projection, and risk analysis to evaluate agricultural real estate and farm and ranch management decision-making.

PRESS SUMMARY

Agriculture faces increasing pressure to remain both profitable and sustainable. This study introduces a farm financial simulation model designed to evaluate how factors such as land tenure, leverage, government programs, and market fluctuations influence farm financial stability over time. By modeling operations of 400, 800, and 1,200 acres, the research generates income statements, balance sheets, cash flow reports, and financial ratios to assess resilience under real-world conditions. This work provides agricultural professionals, policymakers, and stakeholders with a decision-making tool to balance long-term profitability with sustainability in today's evolving agricultural economy. The research is relevant to anyone interested in agricultural risk management, profitability, ownership strategies, or related financial considerations.

A Statistical Look into how Common Soccer Metrics Influence Expected Goal Measures in the Professional Game

Tristan Rumsey[a] *& Shaha Alam Patwary**[b]

[a] *Eli Lilly and Company, Indianapolis, IN and Department of Mathematical Sciences, Butler University, Indianapolis, IN*
[b] *Department of Mathematical Sciences, Butler University, Indianapolis, IN*

https://doi.org/10.33697/ajur.2025.161

Student: tristangrumsey@gmail.com
*Mentor and Corresponding Author: mpatwary@butler.edu**

ABSTRACT
The advent of sports analytics has ignited a fervor across all sporting disciplines, particularly soccer, where clubs are sprinting to harness vast data reserves to elevate team performance, spearhead effective marketing endeavors, and bolster financial gains crucial for club expansion. Much like Billy Beane's transformative "Moneyball" approach, soccer clubs are in pursuit of innovative strategies to transcend financial limitations and achieve triumph. In soccer, where goals are scarce commodities, heightened offensive efficacy becomes imperative. Presently, one metric stands out as pivotal in gauging a team's goal-scoring success: expected goals (xG). This metric quantifies the likelihood of a given shot or opportunity culminating in a goal, making it a linchpin in a team's offensive strategy. Maximizing expected goals becomes paramount for teams aiming to capitalize on limited scoring opportunities during matches. Crucially, the first step in reshaping tactical approaches hinges on identifying the most influential variables in predicting expected goals. This study employs an array of machine learning methodologies, including Ridge, Lasso, Elastic Net, and Group Lasso models. The objective is to unveil the key predictor variables that significantly impact team (offensive) performance, often delineating the thin line between championship glory and defeat. With the aim of predicting xG, this research also incorporates modified bootstrap techniques to compute prediction intervals for the regularized machine learning models. By delving into the intricate fabric of soccer analytics, this study seeks to empower clubs with actionable insights, fostering a new era of strategy and competitive edge on the field.

KEYWORDS
Soccer analytics; Expected goals; Managerial strategy; Statistical and machine learning methods; Bootstrap method; Prediction interval.

INTRODUCTION

In 2007, Liverpool Football Club faced Associazione Calcio (AC) Milan in the Champions League Final – a rematch of their dramatic 2005 encounter. While the final scoreline (2–1 to Milan) reflected a typical result, it did not capture the broader story. Milan had long prepared for this moment through its innovative "Milan Lab," where doctors and performance analysts examined players' physiological and biomechanical data such as jumping ability, heart rate, muscle weakness, and eye movement. The lab claimed that jump metrics alone could predict injuries with 70% accuracy.[1] This represented one of the earliest instances of using data to optimize player health and performance in soccer. The explosion of sports analytics began with the Oakland Athletics' revolutionary "Moneyball" strategy in the 2002 baseball season.[2] The Athletics' success demonstrated how data–driven strategies could allow underfunded teams to compete with wealthier clubs. Since then, analytics has evolved into a multibillion-dollar global industry, valued at more than $2.5 billion in 2022 and projected for rapid growth.[3] Despite initial skepticism from traditionalists such as Wilbon,[4] Rose,[5] and Cowher,[6] analytics now shape nearly every modern sport.

Compared to American sports, soccer was slower to embrace analytics.[7] Historically, managerial decisions were driven by instinct and short–term performance.[8] However, clubs now collect vast data on match statistics, training sessions, and player fitness. Jean–Pierre Meersseman of AC Milan likened analytics to a car dashboard–helpful information that "makes driving easier."[1] This "dashboard" of data empowers coaches, players, and executives to make informed, evidence-based decisions. Among modern soccer metrics, none has been more influential than expected goals (xG).[1] Expected goals quantify "the probability of a shot resulting in a goal."[9] For example, since 78% of penalties in professional soccer are scored,[10] each penalty carries an xG value of 0.78 rather than a binary outcome of 0 or 1. Summing these probabilities across all shots in a match gives a team's total xG, a key measure for evaluating offensive performance and tactical efficiency. This study focuses on identifying the variables most associated with a team's expected goal total.

Because soccer features few scoring opportunities, understanding which metrics most affect expected goals is crucial. A team's xG reveals what truly occurred on the field–far better than the final score alone.[11] A single goal can define a season; optimizing player performance and team tactics to maximize xG can mean the difference between winning and falling short. Recent advances in soccer analytics aim to model how both on-ball and off-ball actions contribute to success. The central question is how "any action changes the likelihood of scoring."[9] Statsbomb's models consider player positions, shot type, and shot quality, while Perl et al.[12] highlight pattern recognition and machine learning as emerging research frontiers. Data from cameras, passes, dribbles, and positional tracking have expanded the analytical scope of the game.[13]

Before developing this study's models, several existing approaches to goal prediction were examined. Sánchez Gálvez et al.[14] used Logistic Regression, Naive Bayes, Decision Trees, and SVMs to predict match outcomes. Decroos & Davis[15] applied a Generalized Additive Model (GAM) to estimate the probability of imminent goals. Inan[16] used Poisson regression to model goal frequency across teams, and Liu et al.[17] implemented transfer and vision learning methods with an Inflated 3D Network (I3D) model to predict goal likelihoods. These frameworks collectively informed the model selection process in this research. A substantial body of work also validates expected goals as a key metric for performance analysis. Historically, narratives in soccer were driven by outcomes rather than quality of play, but xG reframes performance by evaluating shot quality rather than results.[11] Expected goals capture the fairness of performance–highlighting when teams win by luck or fail despite dominance. Because xG incorporates nearly every event leading to a shot, it provides a nuanced view of team and player performance.

In a study of 5,020 matches, 1,366 matches had xG values matching the final score, and 3,443 matched within a one-goal margin.[8] This strong correlation underscores xG's reliability as a predictor of performance and its growing influence in the sport. Statsbomb's analyses further show that shooting from central areas, favoring foot shots over headers, limiting crosses, and improving finishing are all correlated with higher xG.[9] Over an entire season, these insights become powerful indicators of long-term team strength.[18] Opta, another leading analytics firm, uses an XGBoost-based model that incorporates contextual data such as distance to goal, shooting angle, goalkeeper position, defender pressure, and play type (e.g., fast break, set piece).[19] While both Opta and Statsbomb measure expected goals, their models differ in computation–Opta's inclusion of goalkeeper positioning being one example. Consequently, reported xG values can vary slightly depending on the data provider.

Still, expected goals alone cannot explain match outcomes. Shots account for only about 1.5% of all match events,[20] meaning that actions such as passing, dribbling, tackling, and transitions must also be included to model performance accurately. Thus, this study incorporates multiple offensive and defensive variables, extending beyond shot-based data to capture the full complexity of play. The growing adoption of analytics is transforming soccer management. Billy Beane, the pioneer of "Moneyball," now advises Dutch club AZ Alkmaar, which uses data–driven methods to compete successfully against much wealthier teams.[21] Similar success stories, including Brentford FC in the English Premier League, illustrate how analytics empower smaller clubs to achieve sustainable success through strategic efficiency.

This research explores the intersection of statistics, machine learning, and soccer, emphasizing how data–driven insights can improve decision-making. Players' livelihoods and team success depend on optimizing performance, and understanding which variables drive goal-scoring opportunities has both strategic and financial implications. This study contributes to that understanding by identifying the most significant predictors of expected goals, providing actionable insights for coaches, analysts, and executives.

This manuscript is organized into six sections. The **Introduction** outlines the background, motivation, and literature supporting the study. The **Materials and Procedures** section describes the dataset, variable definitions, data cleaning, and exploratory data analysis (EDA). The **Methods and Procedures** section details the statistical and machine learning methodologies used. The **Results** section presents model performance metrics, accuracy measures, and predictive comparisons. The **Discussion** interprets the key findings and implications, while the **Conclusion** provides a synthesis of insights and recommendations for future research.

MATERIALS AND PROCEDURES

The Dataset

The dataset used for this study worked was manually created with data from Football Reference (known as *FBref.com*). This site was created by Sports Reference to document many statistics in professional soccer matches around the world, dating back to 2017, for many top clubs in their various competitions. These competitions may include domestic leagues, domestic cups, intra-European/continental, or even friendly (exhibition) matches. The data stored on *FBref.com* is collected by Opta, who captures and shares real-time sports data with other companies, and professional teams.

This analysis studies the data for Arsenal Football Club, a well-known team in the English Premier League, over the past four league seasons, beginning with the 2019-2020 season. The English Premier League is widely regarded as the most competitive league in the world, and is currently recognized as such, as the Union of European Football Association (UEFA) has ranked it as having the highest league coefficient ranking. Throughout the duration of these four seasons, there have been a number of domestic opponents that Arsenal has faced, including the following clubs: Aston Villa, Bournemouth, Brentford, Brighton & Hove Albion, Burnley, Chelsea, Crystal Palace, Everton, Fulham, Leeds United, Leicester City, Liverpool, Manchester City, Manchester United, Newcastle United, Norwich City, Nottingham Forest, Sheffield United, Southampton, Tottenham Hotspur, Watford, West Bromwich Albion, West Ham United, and Wolverhampton Wanderers.

The complete dataset contains a total of twenty one variables (including the response variable), with a total of 152 records, accounting for thirty eight league matches per season, spanning four seasons. It is important to note that, while Arsenal plays many non-league matches throughout the course of a season, the number and frequency of additional matches is variable. Therefore, to keep each season consistent with the others, this study only analyzes data from the league-specific matches. Many of these records are quantitative data types, but qualitative data will also be present. The dataset was split into train and test datasets. The train dataset includes data from the first three seasons of analysis, including the 2019-2020 (1), 2020-2021 (2) and the 2021-2022 (3) seasons, and the test dataset includes data for the most recently completed season at the time of analysis, 2022-2023 (4). The train dataset will be used to train a model that will then be used to predict values for comparison to the test dataset to evaluate the performance of the adopted model.

Response and Predictor Variables

As the main objective of this research is to identify the metrics that most influence a team's total number of expected goals (xG) in a match, the variable *xG* will be focused on as a response. The response variable is continuous in nature. Additionally, the "Match" variable serves as a unique identifier for each observation. The nineteen remaining predictor variables for expected goals include:

Predictor Variable	Description
Season	Label representing the season that each record corresponds to (numbered 1 through 4)
Team formation	Formation used to start the game
Possession	Percentage of the game that a team has control of the ball
Passing accuracy	Number of passes completed divided by total number of passes
Short length passing accuracy	Accuracy for passes between 5 and 15 yards
Medium length passing accuracy	Accuracy for passes between 15 and 30 yards
Long length passing accuracy	Accuracy for passes greater than 30 yards
Dribble success rate	Success rate of taking on a defender while dribbling
Percentage of shots on target	Percentage of shots taken that would be a goal, if no goalkeeper were to be present
Opponent fouls	Number of fouls committed by the opposing team
Offsides	Number of times a team gets penalized for having an attacking player behind the last defender
Recoveries	An action that ends possession for the opponent and begins possession for the other team
Touches in the attacking third	Touches in the attacking team's final third of the field
Touches in the attacking penalty area	Touches in the attacking team's penalty box
Tackle win percentage	Proportion of tackles where the tackler's team won possession
Crosses	Medium-long range pass angled toward the center of the field near the goal
Corner kicks	Place kick taken by the attacking team, from the corner nearest the goal
Interceptions	Defensive player intercepts the ball from its intended target
Home/away status	Team's status of being home or away in a match

Table 1. Table of Predictor Variables and Description

In the raw dataset, there existed several variables that provided highly repetitive information, and required consideration for removal. A further discussion of those several variables is included in the following section, which discusses pre-processing steps taken. There were a multitude of pre-processing steps that were necessary prior to beginning analysis of the raw dataset.

Data Pre-Processing

Removal of Redundant Variables

Several of the predictor variables in the raw dataset are similar in nature and required removal due to their repetitive nature. One variable that was removed was passing accuracy. There was a total of four distinct variables that related to passing accuracy. Three of these four variables included short length, medium length, and long length passing accuracy. Short length passing accuracy is the accuracy of passes that are between 5 and 15 yards, whereas medium length is the success rate of passes between 15 and 30 yards, and long length pass completion is for passes that are greater than 30 yards. Because a single pass cannot be represented by multiple of these three pass length types, these three variables were all justifiably independent from one another. However, because the passing accuracy variable is a cumulative average of all a team's passing success (short, medium, and long) for a given match, it was best to consider removing the passing accuracy variable.

Additionally, the raw dataset included two variables relating to number of touches: touches in the opponent's final third of the field, and touches in the opponent's penalty area. In further thinking, it became clear that these variables were inherently correlated. If the attacking team happened to make a touch in the opponent's penalty box, there would have been a very strong likelihood that they had already had one or more touches in the opponent's final third of the field. Apart from long crosses or through balls, there are a limited number of ways where the team could ultimately attain a touch in the penalty area without already going through the attacking third of the field. Given this, it was de-

cided to remove the predictor variable of number of touches in the opponent's final third, and to keep touches in the penalty area.

Conversion to Factor Variables

There were several qualitative variables in the raw dataset, that required conversion to factor variables, prior to beginning modeling and analysis. These variables included season, formation, and home/away status. Each season was converted to be a number, labeled 1 through 4, to numerically represent each of the four different league seasons. Similarly, over the course of the four total seasons of data, Arsenal utilized a total of 11 total formations, that were each given a label. Finally, the home or away status variable had two possible values (home or away), and additionally required factor conversion. Upon the completion of this process, these three predictor variables were sufficient to be included in modeling.

Bivariate Vizualization and Analysis of Data

Below are scatterplots for each quantitative variable in the dataset, versus the response variable, expected goals (xG), as well as a table showing the correlation (r) value for each predictor variable versus the response variable. These figures (figure 1–4) and table 2 show the relationship between each numeric predictor variable and the response variable, giving us insight into how variables may or may not be correlated to expected goal values.

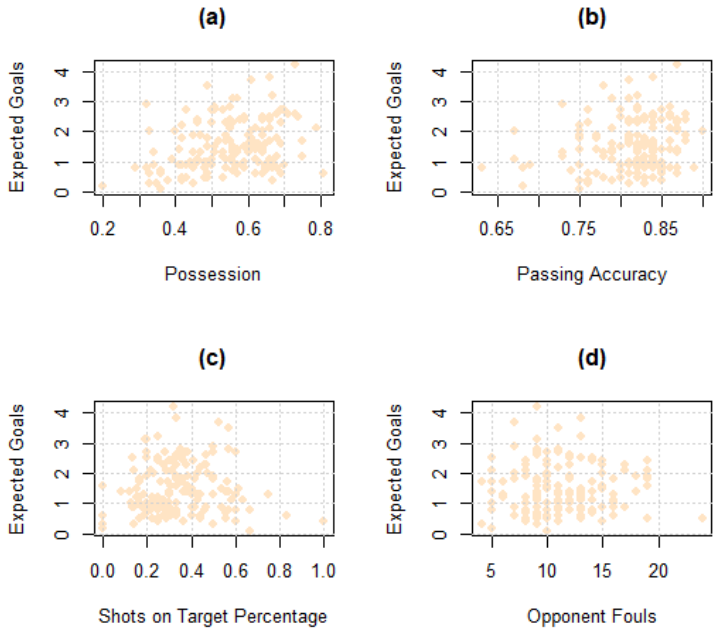


Figure 1. Scatterplots of Possession (a), Passing Accuracy (b), Percentage of Shots on Target (c), and Opponent Fouls (d) vs Expected Goals

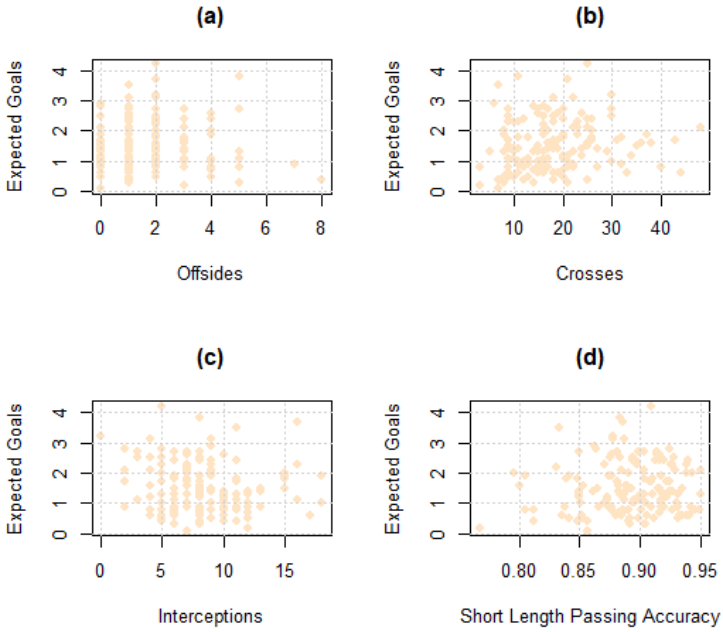


Figure 2. Scatterplots of Offsides (a), Crosses (b), Interceptions (c), and Short Length Passing Accuracy (d) vs Expected Goals

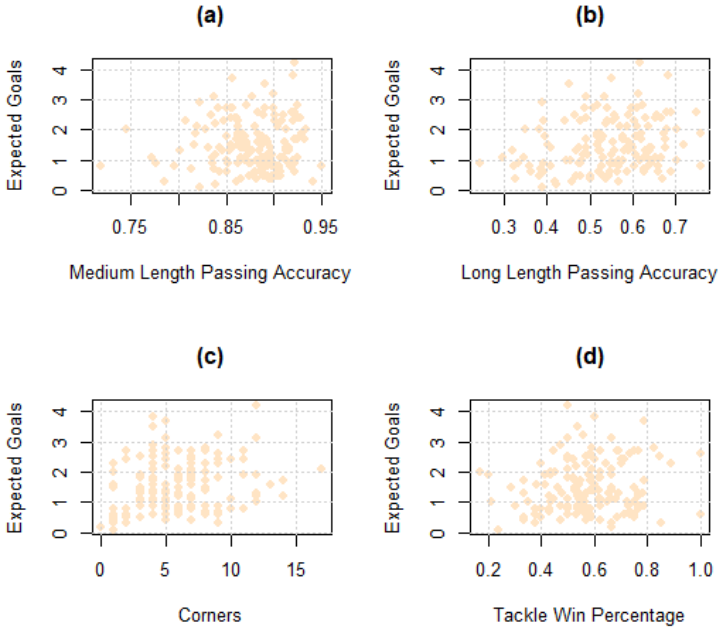


Figure 3. Scatterplots of Medium Length Passing Accuracy (a), Long Length Passing Accuracy (b), Corners (c), and Tackle Win Percentage (d) vs Expected Goals

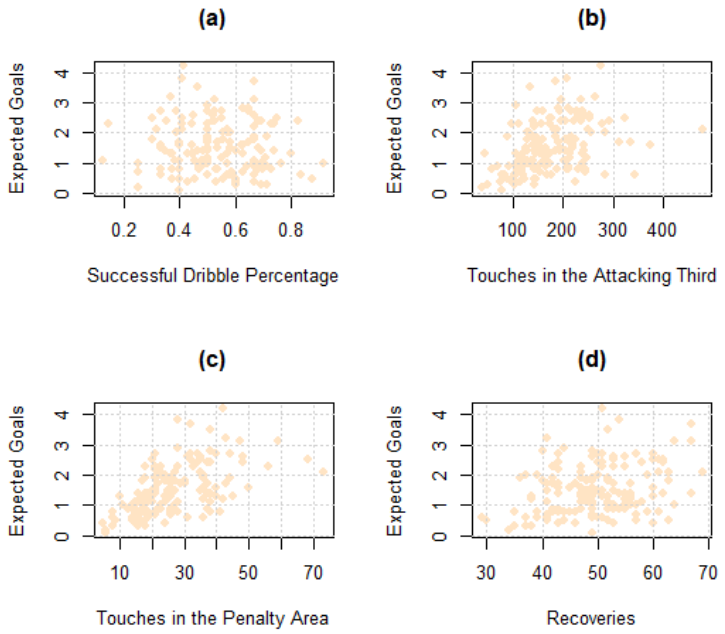


Figure 4. Scatterplots of Successful Dribble Percentage (a), Touches in the Attacking Third (b), Touches in the Penalty Area (c), and Recoveries (d) vs Expected Goals

Predictor Variable	Product-moment Correlation Coefficient (*r*)
Possession	0.31
Passing Accuracy	0.16
Percentage of Shots on Target	0.02
Opponent Fouls	-0.02
Offsides	-0.01
Crosses	0.15
Interceptions	-0.11
Short Length Passing Accuracy	0.05
Medium Length Passing Accuracy	0.06
Long Length Passing Accuracy	0.22
Corners	0.23
Percentage of Tackles Won	0.12
Percentage of Successful Dribbles	-0.06
Touches in the Attacking Third	0.40
Touches in the Opponent's Penalty Box	0.59
Recoveries	0.27

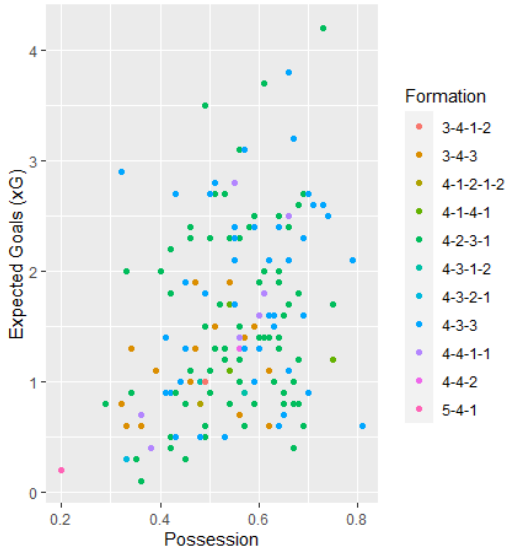
Table 2. Table of Pearsonian Product-moment Correlation Coefficient (r) for Predictor Variable vs Expected Goals (xG)

As seen in the scatterplots and table above, there are several variables that have some linear correlation with expected goals. Possession, long length passing accuracy, corners, touches in the attacking third and penalty box, and recoveries are all variables that had some correlation to the response variable. These correlations were studied in greater detail in the Results and Analysis section of this study, discussing multicollinearity issues in the dataset.

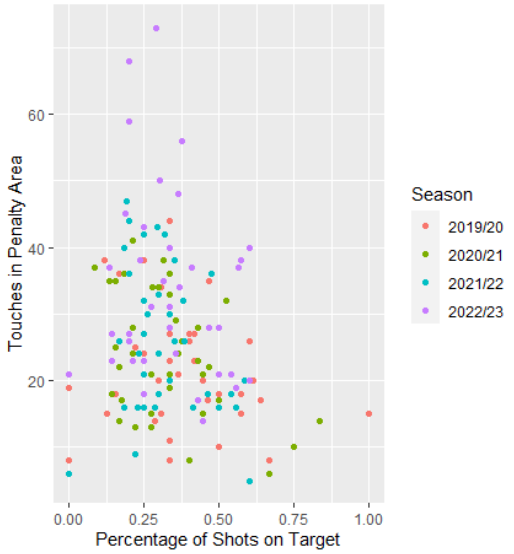
Additional Data Visualizations

From analysis of the scatterplots, it is seen that several variables, including possession, passing accuracy, percentage of shots on target, crosses, touches in the attacking third and penalty box, and recoveries were all correlated with expected

goals. To explore further graphical trends among these variables and several of the factor variables, several additional plots (figure 5–6) were created to highlight other interesting findings in the pre-processing stage. These plots, as well as a brief discussion for the findings from each, are given below.

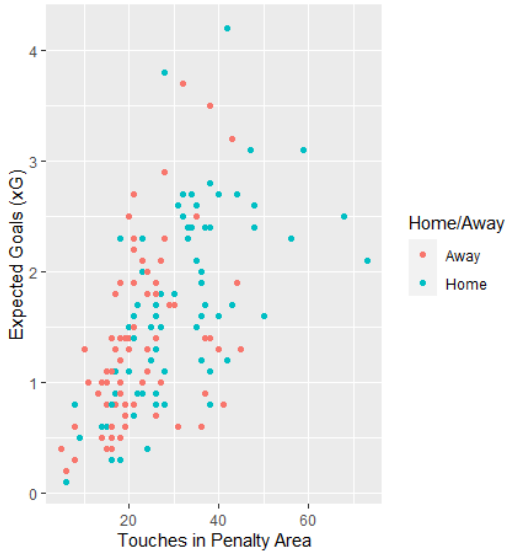


(a) Scatterplot of Possession vs xG, by Formation

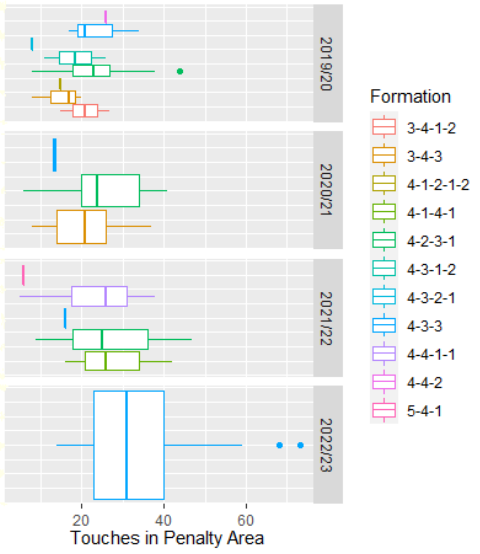


(b) Scatterplot of Shots on Target Percentage vs Touches in the Penalty Area, by Season

Figure 5. Plots Analyzing Possession, Shots on Target Percentage, and Touches in the Penalty Area



(a) Scatterplot of Touches in the Penalty Area vs xG, by Location



(b) Boxplot of Touches in the Penalty Area, by Formation and Season

Figure 6. Plots Analyzing Touches in the Penalty Area

In the first plot, **Figure 5 (a)**, it is clear that team possession was much higher for particular formations, including 4-3-3, 4-2-3-1, and 4-4-1-1. This implies that Arsenal achieved greater success in holding onto the ball, but does not yet confirm if possession is significant in predicting expected goals. The second plot, **Figure 5 (b)**, shows that there is no clear relationship between percentage of shots on target and number of touches in the penalty area. Additionally, percentage of shots on target generally varies by season, whereas touches in the penalty area is clearly higher in the fourth season. Knowing that Arsenal had performed the best in the most recent (fourth) season, it may have been expected that the team would have a higher percentage of shots on target, and also a higher number of touches in the penalty area. It was surprising to see that percentage of shots on target did not vary much throughout different seasons.

The next two plots focus more specifically on the relationship of touches in the penalty area and expected goals. Touches in the penalty area was highly correlated with expected goals based on its scatterplot, and it was of interest to explore further. First, in **Figure 6 (a)**, it can be seen that there tends to be a much greater number of total touches in the penalty area while playing at home. This result would be expected, as home field advantage does generally exist in most sports, but it was interesting to see just how much more pressure Arsenal was placing on their opponents while playing in North London. Finally, in **Figure 6 (b)**, the total number of touches in the penalty area was explored by season and by formation. It was peculiar to see how the number of formations used throughout the season continually decreased over time. Again, knowing that Arsenal performed well in the most recent season where only one formation (4-3-3) was used, it appears that experimenting with fewer formations may be a factor in improving team performance. This may be because certain formations are more comfortable and a natural fit for the players, and trying a new formation may throw off team chemistry and comfort.

METHODOLOGY

Methodology

This section details the several different regression modeling techniques used throughout this analysis. Due to the failure of several model assumptions, it was necessary to explore more complex techniques than multiple linear regression, to identify which predictor variables are most important in predicting expected goals. Each of these different models are described in the following subsections, including their formulas and purpose.

Multiple Linear Regression (MLR) Model

Prior to beginning a deep analysis, preliminary model assumption checking was required. This process began by creating a baseline multiple linear regression model that incorporated all original nineteen predictor variables. The initial multiple linear regression model was as follows,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \textbf{Equation 1.}$$

where $\mathbf{Y}$ is a $n \times 1$ vector of expected goals, $\mathbf{X}$ is a $n \times p$ design matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of random error components.[22, 23] Here, $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$. Therefore, $\mathbf{Y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$.

Residual Diagnostics and Multicollinearity Analysis

Using this initial multiple linear regression model, the key assumptions for linear regression were tested. The five key assumptions are for linearity, homoscedasticity, independence, normality, and multicollinearity. The first and second assumptions lie in checking the residual plot. The residual plot must show a random distribution of residuals, and also show a distinct horizontal band shape, which suggests that the variances of the errors are equal. The next model assumption that required checking was for a linear Q-Q plot, which determines how normal the initial dataset is. Finally, the remaining two plots focus on the independence of assumptions.[24]

Additionally, the initial multiple linear regression model is used to identify the Generalized Variance Inflation Factor (GVIF) values for each predictor variable. A GVIF value is a generalized version of the Variance Inflation Factor (VIF)

value. A VIF represents the amount of variance that is inflated, for each individual predictor variable in a multiple linear regression model. These coefficients have larger values when multicollinearity exists in the data. A VIF value of 1 means that the predictor variable in question has no correlation to any other predictor in the model. In contrast, a VIF value greater than 10 implies that the variable exhibits signs of multicollinearity and a correlation to at least one other predictor variable, necessitating correction.[25] A GVIF, in contrast, additionally accounts for variables that have greater than 1 degree of freedom, such as polynomials, or categorical variables with more than 2 levels. GVIF applies the appropriate change needed to properly analyze if a particular variable has issues relating to multicollinearity.[26]

Variance Stabilizing Transformation

Transforming the response, and/or the predictor variables, can play a great impact by improving model fit, and to achieve a standard closer to normality. To identify the optimal transformation (or lack thereof), the Box-Cox transformation method is often used. The Box-Cox method is a commonly used statistical technique that specifically changes the response variable, to resemble a normal distribution more accurately. This process is accomplished using a λ value, which represents the most optimal exponent to apply to the response variable data. This λ value can range from -2 to 2, with possible values including -2, -1, -0.5, 0, 0.5, 1, or 2; each value is associated with a transformation, including inverses, square roots, natural logs, or exponents. Additionally, a λ value of 1 results in a conclusion that no transformation is the best fit for the data.[27] The Box-Cox method is critical in determining if a transformation of the response variable is the best option to improve overall model performance. This decision and respective transformation must be made prior to modeling. The results of the Box-Cox test will be discussed in the Analysis section of this paper.

Likelihood Ratio Test

The likelihood ratio test is a form of hypothesis testing that enables one to choose the best model between two possible options, based on the ratio of their likelihoods.[28] The two models being compared are usually split, with one model being simple, and the other being complex. In this study, the simple model was the linear regression model without passing accuracy and touches in the attacking third (as both were found to have natural correlation to other predictors), and the complex model was the same model, but with the applied square root transformation, as was found optimal in the Box-Cox procedure. The formula for the likelihood ratio test represents the ratio between the log likelihood (L) of the simpler model and the more complex model. The result of this equation yields a chi-square value to use for testing the null hypothesis that the simpler model is the best fit, versus the alternative that the complex model is a better fit. The degrees of freedom for the test is equivalent to the difference in number of total parameters for the two individual models.[29] The likelihood ratio test procedure is important in determining if model complexities, such as additional variables and transformations are justifiable to implement. The results of the likelihood ratio test will be discussed further in the Results and Analysis section of the paper.

R^2 and AIC

The previously mentioned statistical and machine learning methods are all potentially applicable to the dataset for Arsenal Football Club, which has a sparse sample size and a large number of predictor variables. Machine learning is critical in finding conclusions from a wealth of data. Each of these methods (Ridge, Lasso, Elastic Net, and Group Lasso) create a model, and this study analyzes and compares the effectiveness of each model in predicting a team's total expected goals, by using adjusted R^2, Akaike Information Criterion (AIC), and other relevant criterion values. The adjusted R^2 value is different from the traditional R^2 value, as it more properly adjusts for a higher number of predictor variables in the model.[24] The adjusted R^2 value adequately penalizes for the number of predictor variables present, as they can implement bias. This method involves dividing the sum of the squares of the residuals and the totals, by their respective degrees of freedom, where n is the total number of observations in the dataset, and k represents the number of estimated parameters in the model, which includes a total of p variables and the intercept, β_0.

Another common statistic used to determine overall model success is AIC. AIC represents how well a model fits the data, for future prediction. Lower AIC values represent better model fit and are considered in conjunction with ad-

justed R^2 to determine which model is the most appropriate for analysis, as well as the appropriate variables to keep in the model,[30] ultimately to determine the goodness of fit of the model in representing the dataset.[31]

All these procedures, as well as pre-processing and analysis steps, have been completed in using the statistical programming language R, which has sufficiently allowed for the possibility of making conclusions from this study. The AIC function in R Studio is not compatible with certain advanced regression models, including those that are generalized linear models (GLM), such as Ridge, Lasso, Elastic Net and Group Lasso. However, AIC is still a highly effective and useful tool to compare initial models created for the dataset.

Variable Selection Methods

Stepwise Regression Methods

Stepwise regression is a procedure that builds a regression model from the given predictor variables and either keeps or removes predictors until they are no longer significant enough to consider.[32] There are two unique types of stepwise regression: forward and backward. A forward stepwise selection involves building a greedy algorithm, with a model beginning with only an intercept (no predictor variables), and adding predictors until they are no longer necessary to predict for expected goals. The model output ultimately performs dimension reduction, by never adding the variables that do not play a significant enough role into the model. In contrast, backward stepwise regression considers a complete model with all predictor variables included and eliminates non-significant predictors until a final model is chosen to represent the data, only including predictor variables deemed significant enough.[33] For both types of stepwise regression, each variable is evaluated against a particular criteria, most commonly Akaike's Information Criterion (AIC).[30]

Regularized Regression Methods

Ridge Regression

There are several other prominent machine learning procedures that can effectively handle sparse or correlated features in a dataset, including Ridge regression, Least Absolute Shrinkage and Selection Operator (Lasso) regression, Net Elastic regression, and Group Lasso regression. Ridge regression, also known as $\mathcal{L}_2$ regularization, begins by applying a penalty (the sum of the squared coefficients) to account for correlation in the dataset. This works by shrinking coefficients of correlated predictor variables, which helps achieve smaller variance. This is important in restricting the influence of predictors.[34] When determining the penalty, Ridge regression considers all predictor variables in the model, which therefore makes the method most useful in scenarios where there are many predictor variables present, and all have non-zero coefficients.[35–37]

Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) regression (known as $\mathcal{L}_1$ regularization) functions in a similar manner to Ridge regression, but has a key difference in the penalty term. Lasso methods also apply a penalty term, that is instead based on the magnitude of coefficients, rather than sum of squared coefficients. In contrast to Ridge regression, Lasso does not weight the coefficients of every predictor variable, but rather often picks one of the correlated predictor variables and ignores the rest, effectively performing variable selection.[35] It can remove predictors from the model, by shrinking coefficients to exactly zero, unlike Ridge.[38] Again, it is common to use machine learning techniques including Lasso when dealing with larger datasets and correlated variables.

It is common to compare the results of Lasso regression with those of Ridge, to determine which model produces superior results with minimized error, to reduce or eliminate the impact of correlated variables in the model.

Elastic Net Regression

Elastic Net regression is another popular machine learning technique, effectively combining elements of both Ridge and Lasso regression, as another way to perform variable selection and shrink predictor variable coefficients.[39] This method removes groupings of correlated variables, while also keeping several predictors in the model. A series of cross-validation steps are used to determine how much of the penalty is determined from Ridge regression, and how much is determined by Lasso regression. Elastic Net regression is traditionally viewed as a balance of both Ridge and Lasso regression, shrinking the coefficients, possibly to 0, but not as commonly as Lasso regression. Elastic Net regression is especially noteworthy for its ability to handle bias. With multiple variables that are highly correlated, it is possible for Lasso regression to introduce bias,[34] and therefore it is worthwhile to compare the performance of Elastic Net regression to both Ridge and Lasso.

Group Lasso Regression

Group Lasso regression follows a similar variable selection process as the Lasso method, but differs in its choice of important predictors. Lasso chooses several significant features among all the variables, whereas Group Lasso involves the splitting of variables into groups to complete variable reduction. Upon splitting the data into groups, if a variable is deemed important in a group, the Group Lasso procedure will include all variables in that same group in its model. If a group produces no important predictors, all its variables will be excluded from the model.[40–42]

Measures of Prediction Accuracy: MSE, RMSE, and MAPE

In addition to analyzing adjusted R^2 and AIC, models require the calculation of error values when determining overall model adequacy. There are three main measures that are routinely used, including mean squared error (MSE), root mean square error (RMSE), and mean absolute percentage error (MAPE).

MSE squares the distance of data points from the regression line (errors), to ultimately find the average of the set of errors. This procedure is popular, due to its squaring technique, which gives more weight to outlier data points. RMSE calculates the square root of the MSE, which effectively makes the larger values generated by outlier data points smaller, and therefore easier to interpret.[43] Finally, MAPE is a method to measure the accuracy of a predictive model. It is desirable to minimize the MSE, RMSE, and MAPE values to increase model success.[24] Each of these error calculations are important to consider, as they are critical in determining the optimal model with the best prediction accuracy. The formulas for each of these error values are listed below:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \textbf{Equation 2.}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad \textbf{Equation 3.}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \qquad \textbf{Equation 4.}$$

where n is the number of data points in the dataset, Y_i is the actual value of the response variable (square root of expected goals) for a particular data point, and $\hat{Y}_i$ is the predicted value of the response variable for a data point.

Bootstrap Method for Regression Models

In the analysis of these models, the creation of prediction intervals is hugely important to understanding the range that the estimated value of square root expected goals falls within, and the accuracy of model prediction. The creation of prediction intervals for each individual record in the test dataset requires bootstrapping, a procedure that resamples data from one sample in order to randomly generate a distribution. This procedure can be used to estimate standard errors, bias, obtain prediction intervals, and also to test a hypothesis.[44] The process of bootstrapping involves sampling the records in the test dataset (in this study, 38 total records) a selected number of bootstrap resamples, $B = 1000$. Often, B is chosen as a high number, to ensure better sample size in creating a prediction interval, from the enlarged sampling distribution.[45] The advanced regularized regression methods described above, including Ridge, Lasso, Elastic Net, and Group Lasso, and their functions and required packages in R, do not provide prediction intervals for individual records of the test dataset. Therefore, to generate these intervals, it became necessary to increase the sample size through a bootstrapping resampling technique.

To account for these issues, an alternative method was required to create a prediction interval using bootstrap resampling method. Each test record had a unique predicted square root expected goal value, $\hat{y}$, calculated by multiplying the $\mathbf{X}$ matrix with the $\vec{\beta}$ vector. $\mathbf{X}$ contained all values of the predictor variables for each record, which was then multiplied with $\vec{\beta}$, the vector storing the coefficients for each predictor variable for each model. In total, there existed 38 unique fitted $\hat{y}$ values, stored in $\vec{\hat{y}}$, the vector of all fitted values, for each model. After collecting these 38 fitted values, they were then bootstrapped, with 1000 resamples of size 38 being generated from the original 38 data vectors. Therefore, we have calculated 1000 predicted values by averaging out 38 predicted values from each of the bootstrap resamples. After this step, a prediction interval of desired confidence level $(1 - \alpha)$ for the expected number of goals was then able to be created for each of 1000 records. This process is unique from that of traditional bootstrapping, where random sampling is usually the first step of the procedure. However, this study first required collection of all fitted values, so that enough records were present to then randomly sample and create a distribution. The results of these bootstrapping procedures are discussed in the following subsection.

RESULTS

The following section describes the results of the aforementioned modeling procedures, including model assumptions, data transformation, fitted models, selection of most important predictor variables and of the best model to use, to predict expected goals along with prediction intervals.

Model Assumption and Multicollinearity Checking

As discussed earlier, inspection of preliminary model assumptions is required before using the model for inferential procedure. Upon fitting this multiple linear regression model of expected goals (xG) on the selected predictors, each key assumption referenced in Methodology section was checked. Four of the five assumptions of linear regression can be checked using the first three plots provided in figure 7, below, including (a) residual plot of the raw dataset versus the model's fitted values, (b) Q-Q plot, (c) scale-location plot, and (d) residual versus leverage plot.

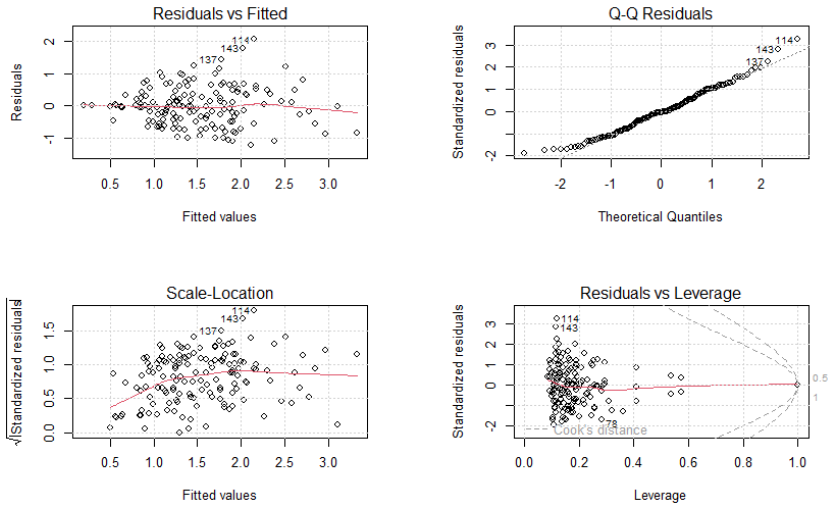


Figure 7. Plots for Simple Linear Regression Model Assumption Checking, Including Residual vs Fitted Values (a), Normal Q-Q (b), Scale-Location (c), and Residuals vs Leverage (d)

Since figure 7 (a) is not showing any specific direction or trend and the red line is very much horizontal, so, linearity and independence of observations assumptions seem reasonable. From figure 7 (b), the points are very close to the dotted straight line except a few points, indicating the normality assumption of error is reasonable. Finally, figure 7 (c) produces the red line as a curve and it has some trend of increasing error variance. Thus, a variance stabilizing transformation of the response variable is necessary for drawing inferential decisions. The multiple linear regression model is as follows:

$$\hat{xG} = 1.164 + 0.149(Season\ 2) + 0.215(Season\ 3) + 0.010(Season\ 4) - 0.424(Formation\ 3-4-3) - 0.965(Formation\ 4-1-2-1-2) - 0.352(Formation\ 4-1-4-1) - 0.277(Formation\ 4-2-3-1) - 0.258(Formation\ 4-3-1-2) - 0.378(Formation\ 4-3-2-1) - 0.220(Formation\ 4-3-3) - 0.263(Formation\ 4-4-1-1) + 0.007(Formation\ 4-4-2) - 0.786(Formation\ 5-4-1) + 1.017(Possession) + 0.874(Passing\ Accuracy) + 0.661(Percentage\ of\ Shots\ on\ Target) - 0.005(Opponent\ Fouls) + 0.048(Offsides) - 0.029(Crosses)^{*} + 0.024(Interceptions) - 1.167(Short\ Length\ Passing\ Accuracy) - 2.099(Medium\ Length\ Passing\ Accuracy) + 0.535(Long\ Length\ Passing\ Accuracy) - 0.011(Corners) + 0.649(Tackle\ Win\ Percentage) + 0.310(Successful\ Dribble\ Percentage) - 0.001(Touches\ in\ the\ Attacking\ Third) + 0.055(Touches\ in\ the\ Penalty\ Area)^{***} + 0.003(Recoveries) - 0.042(Home/Away\ H)$$

Additionally, GVIF values were also analyzed for assessing the last assumption, multicollinearity. Below is a table (table 3) documenting each predictor variable's respective GVIF values and degrees of freedom, in the multiple linear regression model. The predictor variables with GVIF values greater than 10 are considered muticollinear in practice.[24] However, it was plausible to leave all variables in the original multiple linear regression model, knowing that the machine learning algorithms used in modeling would perform variable reduction, and remove these problematic predictor variables at a later time.

Predictor Variable	GVIF	Degrees of Freedom
Season	17.528462	3
Formation	39.744686	10
Possession	5.009569	1
Passing Accuracy	19.563667	1
Percentage of Shots on Target	1.479036	1
Opponent Fouls	1.317865	1
Offsides	1.319266	1
Crosses	4.389041	1
Interceptions	1.607953	1
Short Length Passing Accuracy	4.742041	1
Medium Length Passing Accuracy	6.222260	1
Long Length Passing Accuracy	3.069269	1
Corners	2.914998	1
Percentage of Tackles Won	1.218060	1
Percentage of Successful Dribbles	1.261656	1
Touches in the Attacking Third	7.410396	1
Touches in the Opponent's Penalty Box	3.71142	1
Recoveries	1.601722	1
Home or Away Status	1.379950	1

Table 3. Table of GVIF Values and Degrees of Freedom for Each Predictor Variable

Variance Stabilizing Transformation of Response Variance

To account for the violations of the model assumptions described above, it was necessary to explore possible transformations of the response variable in the dataset, expected goals (xG). As mentioned in the "Variance Stabilizing Transformation" subsection in the Methodology section, the Box-Cox method is a common technique to use to identify the type of data transformation when either of the linearity, normality, and constant variance assumptions about the fitted model is questionable. The Box-Cox procedure allows us to produce a plot of transformation parameter (λ) on horizontal axis and value of likelihood function of the model associated with λ value on vertical axis. The plot for the original multiple linear regression model is shown below, where the confidence interval bounds of the optimal λ value intercept 0.5, but do not intercept values of $\lambda = 0$, nor $\lambda = 1$.

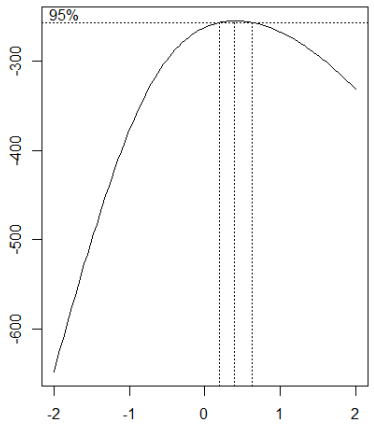


Figure 8. Box-Cox Plot to Identify Type of Variable Transformation

This plot implies that a λ value of 0.5 is the best fit, and therefore the response variable requires a square root transformation to improve normality. Going forward, future modeling would be completed using a transformed response variable: the square root of the expected goal value. The following multiple linear regression model, with response variable transformation, is shown below:

$$\hat{xG_t} = 0.836 + 0.043(Season\ 2) + 0.066(Season\ 3) - 0.003(Season\ 4) - 0.158(Formation\ 3-4-3) - 0.404(Formation\ 4-1-2-1-2) - 0.099(Formation\ 4-1-4-1) - 0.118(Formation\ 4-2-3-1) - 0.093(Formation\ 4-3-1-2) - 0.315(Formation\ 4-3-2-1) - 0.097(Formation\ 4-3-3) - 0.082(Formation\ 4-4-1-1) + 0.048(Formation\ 4-4-2) - 0.507(Formation\ 5-4-1) + 0.446(Possession) - 0.043(Passing\ Accuracy) + 0.218(Percentage\ of\ Shots\ on\ Target) - 0.001(Opponent\ Fouls) + 0.019(Offsides) - 0.011(Crosses)^{*} + 0.013(Interceptions) - 0.085(Short\ Length\ Passing\ Accuracy) - 0.706(Medium\ Length\ Passing\ Accuracy) + 0.246(Long\ Length\ Passing\ Accuracy) - 0.005(Corners) + 0.267(Tackle\ Win\ Percentage) + 0.124(Successful\ Dribble\ Percentage) - 0.001(Touches\ in\ the\ Attacking\ Third) + 0.023(Touches\ in\ the\ Penalty\ Area)^{***} + 0.001(Recoveries) - 0.037(Home/Away\ H)$$

As was done in the previous section, each of the five assumptions must be checked, following this transformation of the response variable. It can be seen that, while there is not perfect normality, there also is no distinct funnel shape, indicating that the overall fit of the model on square-root transformed response (xG) might be more reasonable than the model with the original xG. The same four plots are shown below, for the purpose of comparison. These plots are providing indication of reasonable model assumptions. Using this transformed response variable will be the basis of subsequent analysis in the coming sections.

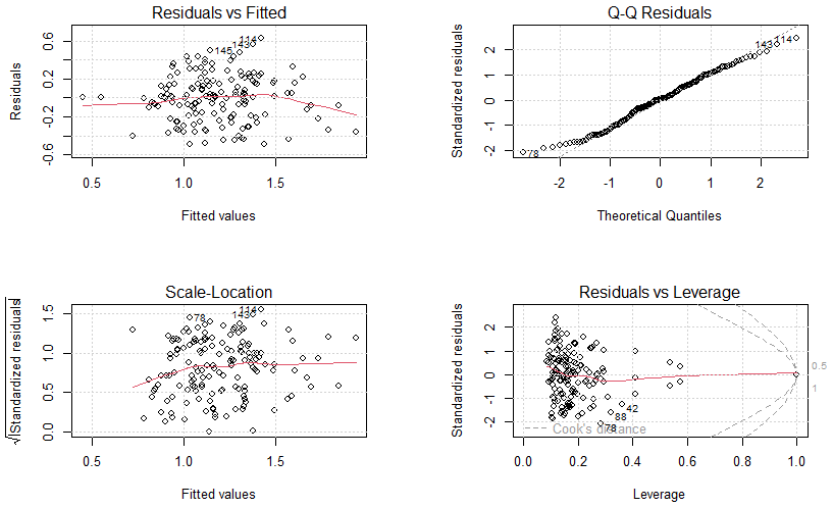


Figure 9. Plots for Model Assumption Checking with the Transformed Response Variable, Including Residual vs Fitted Values (a), Normal Q-Q (b), Scale-Location (c), and Residuals vs Leverage (d)

Results of the Likelihood Ratio Test

The likelihood ratio test is a key step to determine the validity of adding complexities in terms of the predictor variables to the original regression model. The test was specifically used in this study to confirm the results of the Box-Cox transformation, and to suggest that the more complex, transformed model would be a better fit for the data. Using the `lrtest` function in R, through the `lmtest package`, a likelihood ratio test was completed to confirm the use of a square root transformation. This test compared two models, including the "simple" model with all variables–except for passing accuracy and touches in the attacking third, as both were found to be redundant variables–versus the same model but with the square root transformation applied to the response variable, expected goals. This yielded a chi-square test statistic of 272.25, and a p-value of 2.2×10^{-16}. This p-value is practically 0 which is less than a designated α value of 0.05, which therefore suggests rejecting the null hypothesis, meaning that the two models are significantly different from one another. This conclusion means that the square root transformation is satisfactory and justifiable, in order to improve overall model adequacy. These results were important steps to enhance the validity of the subsequent

modeling that has been performed and that will be documented below, which uses the same square root transformation for the response variable.

Fitted Models Under Different Regression Methods

This subsection details results for each of the previously mentioned regression modeling techniques, and provides the fitted model, adjusted R^2, most important predictor variables chosen, and if applicable, the test statistic and p-value from each model. For the regularization methods, the glmnet() function used to create the models (Ridge, Lasso, Elastic Net, and Group Lasso) in R does not generate p-values or confidence intervals for each predictor variable. This limitation required the use of error values and adjusted R^2 to compare overall model performance.

Stepwise Regression Methods

Backward Stepwise Regression We have fitted the model using backward step-wise regression method. The outputs include each selected variable's estimated parameter (β), standard error, test statistic value (t), and p-value, as well as the overall model's residual standard error, degrees of freedom, multiple and adjusted R^2 values, F statistic, and p-value.

The fitted model for the backward stepwise regression model for the transformed response variable are as follows,

$$\hat{xG_t} = 0.412^{**} + 0.280(Percentage\ of\ Shots\ on\ Target)^{*} - 0.009(Crosses)^{**} + 0.021(Touches\ in\ the\ Penalty\ Area)^{***} + 0.006(Recoveries)^{*}$$

F = 28.17

Regression Degrees of Freedom = 4

Residual Degrees of Freedom = 147

P-value = 2.2e-16

Adjusted R^2 = 0.419

AIC = 30.98115

where $\hat{xG_t}$ represents the predicted (square root) value for expected goals. The stars (*) attached to each coefficient represents the significance at different nominal α levels. One star represents that the parameter is significant at the α = 0.05 level, two stars mean that the parameter is significant at the α = 0.01 level, and three stars mean that the parameter is significant at the α = 0.001 level.

The fitted model produced an adjusted R^2 value of 0.4185, meaning that approximately 41.85% of the total variation in the square root of expected goals is explained by these four predictor variables. Considering the fitted model, all four predictor variables selected through this machine learning technique–percentage of shots on target, crosses, touches in the penalty area, and recoveries–are all individually significant in predicting expected goal values. Specifically, percentage of shots on target, recoveries, and touches in the penalty area had a positive impact on a team's predicted total xG in a match, whereas crosses had a negative impact on predicting a team's total xG. This will be contrasted with the variables deemed significant from each of the subsequent models.

Further, the overall model achieved an F statistic value of 28.17, with a regression degree of freedom equal to four, and a residual degree of freedom equal to 147. This test statistic value resulted in a p-value of 2.2×10^{-16}, implying that, in addition to each individual predictor variable, that the overall model is also significant in predicted expected goals. Finally, the backward stepwise regression model generated an AIC value of 30.98115, which will be used to compare to the model generated by forward stepwise regression.

Forward Stepwise Regression The fitted model for the forward stepwise regression model along with the necessary outputs for the transformed response variable was as follows,

$$\hat{xG_t} = 0.412^{**} + 0.280(Percentage\ of\ Shots\ on\ Target)^{*} - 0.009(Crosses)^{**} + 0.021(Touches\ in\ the\ Penalty\ Area)^{***} + 0.006(Recoveries)^{*}$$

F = 28.17

Regression Degrees of Freedom = 4

Residual Degrees of Freedom = 147

P-value = 2.2e-16

Adjusted R^2 = 0.419

AIC = 30.98115

where $\hat{xG_t}$ again represents the predicted (square root) value for expected goals. This fitted model, the corresponding significance levels, and other measures of model success are the exact same as the values found from completing backward stepwise regression. The adjusted R^2 value represents that 41.85% of the total variation in the square root value of expected goals is explained by this forward stepwise regression model, containing these four variables (percentage of shots on target, crosses, touches in the penalty area, and recoveries), which were again found to be individually significant at the α equals 0.05. Further, the overall fitted model was found to be significant, with the same F statistic of 28.17, and a p-value of 2.2×10^{-16}. Again, each of these four predictor variables, in addition to the overall model, were significant in predicting the square root transformed value of expected goals, with percentage of shots on target, recoveries, and touches in the penalty area having a positive impact on xG, and crosses having a negative impact. The AIC of this model was also equivalent to the backward stepwise model, with a value of 30.98.

Regularized Regression Methods

Ridge Regression Model Ridge regression keeps all predictor variables in the fitted model (including those that were deemed to have a natural correlation to other predictors), but appropriately penalizes the less important predictor variables. The fitted model is as follows,

$$\hat{xG_t} = 0.512 + 0.031(Season)0.002(Formation) + 0.333(Possession) + 0.077(Percentage\ of\ Shots\ on\ Target) + 0.002(Opponent\ Fouls) + 0.002(Offsides)0.004(Crosses) + 0.005(Interceptions) + 0.004(Corners)0.136(Short\ Length\ Passing\ Accuracy)0.234(Medium\ Length\ Passing\ Accuracy) + 0.192(Long\ Length\ Passing\ Accuracy)0.003(Percentage\ of\ Tackles\ Won) - 0.005(Percentage\ of\ Successful\ Dribbles) + 0.012(Touches\ in\ the\ Penalty\ Area) + 0.006(Recoveries)0.001(Home\ or\ Away\ Status)$$

Adjusted R^2 = 0.694

where $\hat{xG_t}$ represents the predicted square root value for expected goals. The coefficients in the fitted model have been penalized where appropriate, so that the variables that are most important in predicting expected goals, including the four variables discussed earlier (percentage of shots on target, crosses, touches in the penalty area, and recoveries), all have a higher weight and importance in calculating the square root value of expected goals.

It is important to note that the regularization methods (ridge, lasso, elastic net, and group lasso) do not use an F test statistic, degrees of freedom, nor AIC values, unlike both backward and forward stepwise selection processes, due to the differing methodologies previously described. However, adjusted R^2 is used across all models to determine overall model success. The ridge regression model achieved an adjusted R^2 value of 0.694, which is a step above the results

achieved by the forward and backward stepwise regression models. Approximately 69.4% of the total variation in the square root of expected goals is explained by this model, which is a quite high percentage, indicating that the model fits the data well.

Lasso Regression Model Lasso regression removes predictor variables that were deemed to not be important in the overall success of the model, by penalizing them and reducing their coefficient to 0. This change is shown in the following fitted model for predicting the square root of expected goals, which contains only several of the nineteen total predictors.

$$\hat{xG}_t = 0.699 + 0.015(Touches\ in\ the\ Penalty\ Area) + 0.002(Recoveries)$$
$$\text{Adjusted } R^2 = 0.693$$

This model indicates that Lasso regression only found two predictor variables to be sufficient for predicting the square root expected goal values. These two variables are touches in the penalty area and recoveries–both of which have been deemed significant in other modeling techniques, such as forward and backward stepwise regression. Both predictors positively impact a team's total expected goal count in a match. The Lasso regression model had an adjusted R^2 value of 0.693, which is just below the adjusted R^2 of the ridge model (0.694).

Elastic Net Regression Model The net elastic regression model serves as a compromise of the ridge and lasso models, with the ability to penalize variables to a coefficient value of 0, but it may keep the coefficient of enough important predictor variables. The following fitted model shows the coefficients generated through elastic net regression.

$$\hat{xG}_t = 0.702 + 0.013(Touches\ in\ the\ Penalty\ Area) + 0.002(Recoveries)$$
$$\text{Adjusted } R^2 = 0.696$$

While this fitted model only kept the same two predictor variables as found in lasso regression, it is important to note the slightly different coefficient values, indicating that each variable was penalized and weighted differently through this method. However, these two variables were again determined to be the most important in predicting the square root value of expected goals, both positively impacting the count of expected goals. The elastic net model had a higher adjusted R^2, informing that the model and its predictor variables chosen account for a greater proportion of the total variation in the transformed response variable.

Group Lasso Regression Model The group lasso model was the final model chosen to represent the dataset, splitting predictors into groups. In R, group lasso is performed using the same glmnet() function, but instead using adding the multinomial parameter to the code, indicating a "grouped" type, instead of the default (no grouping). Similar to the lasso regression model, the group lasso model can penalize insignificant predictor variables to the value of 0, keeping only the predictors that are significant enough. The fitted model below shows the coefficients generated through group lasso regression.

$$\hat{xG}_t = 0.699 + 0.015(Touches\ in\ the\ Penalty\ Area) + 0.002(Recoveries)$$
$$\text{Adjusted } R^2 = 0.693$$

The fitted model has again selected touches in the penalty area and recoveries to be two particularly strong predictors to model the square root value of expected goals. These coefficient values are similar to those generated from the

lasso and elastic net models. This implies that all the models found touches in the penalty area and recoveries to be especially important in prediction of square root expected goals, each positively impacting the expected goal tally. The adjusted R^2 value generated by this model was 0.693. These results are less successful than those of elastic net regression. The group lasso model had slightly higher error values, indicating slightly worse accuracy, and had a lower adjusted R^2 than the Elastic Net model. This informs us, that while all models performed well, the Elastic Net model was marginally better than the other models employed in analysis.

Results and Accuracy Measures of the Fitted Models

The prediction accuracy measures of each of the six non-transformed regression and machine learning models are shown in the table below. It can be seen that many of the models, especially the regularization methods, produced the best accuracy measures. Out of all, the elastic net model performed the best, as it contained both the highest adjusted R^2 value, and produced the lowest error values.

	MSE	RMSE	MAPE
Backward	0.569	0.7546	44.88%
Forward	0.615	0.7843	47.35%
Ridge	0.596	0.7718	45.57%
Lasso	0.620	0.7877	43.58%
Elastic Net	0.616	0.7846	42.71%
Group Lasso	0.620	0.7877	43.59%

Table 4. Prediction Accuracy Measures by Model

Prediction Intervals

Following modeling procedures, it became necessary to produce prediction intervals for the 38 predicted expected goal values from the test dataset. This process yielded a 95% prediction interval for each of the predicted expected goal values, comparing the predicted value to the lower (2.5%) and upper (97.5%) bounds of the interval. To generate the prediction interval bounds for each of the 38 predicted values, bootstrapping was required because of the relatively small size of the number of predicted values, as well as the inability of these machine learning models to generate their own prediction intervals. Each of the 38 records in the test dataset were randomly bootstrapped a total of 1000 times in the creation of a prediction interval, to test if each predicted expected goal total value fell inside or outside of the 95% interval.

This procedure was repeated for each modeling technique–Ridge, Lasso, Elastic Net, and Group Lasso–on the non-transformed response variable. The bounds of the prediction intervals, as well as the predicted values for the test dataset generated by each model, are shown in the tables (table 5–6) below, with a further discussion taking place afterwards.

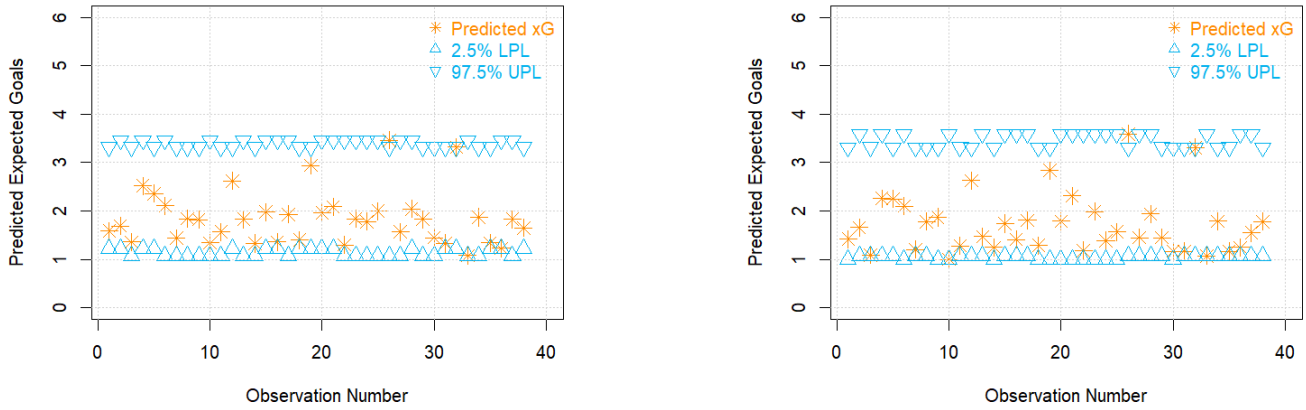
Ridge			Lasso		
Predicted xG	2.5% LPL	97.5% UPL	Predicted xG	2.5% LPL	97.5% UPL
1.59	1.22	3.32	1.42	1.00	3.31
1.69	1.23	3.46	1.66	1.07	3.59
1.36	1.08	3.32	1.09	1.07	3.32
2.52	1.23	3.46	2.26	1.07	3.59
2.36	1.22	3.32	2.25	1.07	3.31
2.11	1.08	3.46	2.10	1.00	3.59
1.44	1.08	3.32	1.20	1.07	3.31
1.84	1.08	3.32	1.78	1.07	3.31
1.82	1.08	3.32	1.87	1.00	3.31
1.34	1.08	3.46	1.00	1.00	3.59
1.57	1.08	3.32	1.28	1.07	3.32
2.62	1.23	3.32	2.63	1.07	3.31
1.83	1.08	3.46	1.47	1.07	3.59
1.33	1.08	3.32	1.25	1.00	3.31
1.98	1.23	3.46	1.74	1.07	3.59
1.36	1.23	3.46	1.41	1.07	3.59
1.93	1.08	3.46	1.81	1.07	3.59
1.40	1.22	3.32	1.29	1.00	3.31
2.94	1.23	3.32	2.84	1.00	3.31
1.96	1.23	3.46	1.79	1.00	3.59
2.09	1.22	3.46	2.31	1.00	3.59
1.30	1.08	3.46	1.18	1.00	3.59
1.84	1.08	3.46	1.99	1.00	3.59
1.78	1.08	3.46	1.39	1.00	3.59
2.00	1.08	3.46	1.57	1.00	3.59
3.46	1.08	3.32	3.59	1.07	3.31
1.57	1.08	3.46	1.45	1.07	3.59
2.04	1.23	3.46	1.94	1.07	3.59
1.83	1.08	3.32	1.45	1.07	3.32
1.45	1.08	3.32	1.17	1.00	3.31
1.33	1.22	3.32	1.17	1.07	3.31
3.32	1.23	3.32	3.31	1.07	3.31
1.08	1.08	3.46	1.07	1.07	3.59
1.87	1.08	3.32	1.79	1.07	3.31
1.34	1.23	3.32	1.17	1.07	3.32
1.23	1.23	3.46	1.25	1.07	3.59
1.83	1.08	3.46	1.55	1.07	3.59
1.65	1.23	3.32	1.78	1.07	3.31

Table 5. Ridge and Lasso Regression to Predict Expected Goal Values and 95% Prediction Intervals, on the Response

Elastic Net			Group Lasso		
Predicted xG	2.5% LPL	97.5% UPL	Predicted xG	2.5% LPL	97.5% UPL
1.42	1.05	3.05	1.42	1.00	3.31
1.61	1.08	3.30	1.66	1.07	3.59
1.13	1.08	3.06	1.09	1.07	3.32
2.14	1.08	3.30	2.26	1.07	3.59
2.12	1.08	3.05	2.25	1.07	3.31
1.98	1.05	3.30	2.10	1.00	3.59
1.22	1.08	3.05	1.20	1.07	3.31
1.72	1.08	3.05	1.78	1.07	3.31
1.78	1.05	3.05	1.87	1.00	3.31
1.05	1.05	3.30	1.00	1.00	3.59
1.30	1.08	3.06	1.00	1.07	3.32
2.44	1.08	3.05	2.63	1.07	3.31
1.50	1.08	3.30	1.47	1.07	3.59
1.25	1.05	3.05	1.25	1.00	3.31
1.72	1.08	3.30	1.75	1.07	3.59
1.39	1.08	3.30	1.41	1.07	3.59
1.74	1.08	3.30	1.81	1.07	3.59
1.29	1.05	3.05	1.29	1.00	3.31
2.67	1.05	3.05	2.84	1.00	3.31
1.74	1.05	3.30	1.79	1.00	3.59
2.16	1.05	3.30	2.31	1.00	3.59
1.19	1.05	3.30	1.18	1.00	3.59
1.87	1.05	3.30	1.99	1.00	3.59
1.40	1.05	3.30	1.39	1.00	3.59
1.54	1.05	3.30	1.57	1.00	3.59
3.30	1.08	3.05	3.59	1.07	3.31
1.45	1.08	3.30	1.45	1.07	3.59
1.89	1.08	3.30	1.94	1.07	3.59
1.44	1.08	3.06	1.45	1.07	3.32
1.21	1.05	3.05	1.17	1.00	3.31
1.17	1.08	3.05	1.17	1.07	3.31
3.05	1.08	3.05	3.31	1.07	3.31
1.08	1.08	3.30	1.07	1.07	3.59
1.74	1.08	3.05	1.79	1.07	3.31
1.19	1.08	3.06	1.17	1.07	3.32
1.25	1.08	3.30	1.25	1.07	3.59
1.52	1.08	3.30	1.55	1.07	3.59
1.69	1.08	3.05	1.78	1.07	3.31

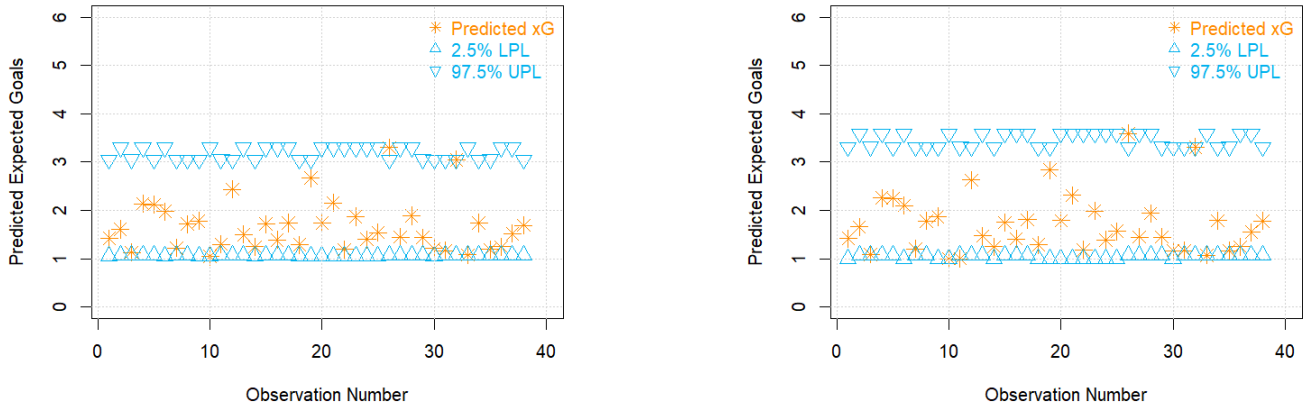
Table 6. Elastic Net and Group Lasso Regression to Predict Expected Goal Values and 95% Prediction Intervals, on the Response

The above tables highlight several interesting features. Each of these advanced regression techniques produced predicted values of expected goals that were relatively similar to one another. This confirms the findings from previous testing, where it was seen that these models were performing with similar success to one another, with a high adjusted R^2 value. Additionally, the vast majority of these predicted values of expected goals fall within the 95% prediction intervals for each test record, as generated through bootstrapping. These results are reflected in the figures (figure 10–11) below, depicting the 95% prediction interval for each modeling technique. The prediction interval for each technique includes its 2.5% lower prediction limit (2.5% LPL) and 97.5% upper prediction limit (97.5% UPL).



(a) Ridge Predicted xG Prediction Interval **(b)** Lasso Predicted xG Prediction Interval

Figure 10. 95% Prediction Intervals for Ridge and Lasso Predictions of Expected Goals



(a) Elastic Net Predicted xG Prediction Interval **(b)** Group Lasso Predicted xG Prediction Interval

Figure 11. 95% Prediction Intervals for Elastic Net and Group Lasso Predictions of Expected Goals

Again, as expected, most of the predicted responses are within the 95% prediction limits. The few predictions that fell outside the bounds of the corresponding prediction interval, however, are deemed to be multivariate outliers. The results of each of these graphs, combined with the results of modeling, show that Ridge, Lasso, Net Elastic, and Group Lasso regression all predicted the expected goal values with acceptable accuracy.

In addition to bootstrapping and creating prediction intervals for each record in the test dataset, each modeling technique required a prediction interval for its overall predicted value of expected goals. A 95% global bootstrap prediction interval (GPL) for each of the advanced regression and machine learning techniques is listed below:

Method	2.5% GPL	97.5% GPL
Ridge	1.658	1.963
Lasso	1.518	1.854
Elastic Net	1.497	1.786
Group Lasso	1.518	1.854

Table 7. 95% global bootstrap prediction interval (GPL)

Each of these 95% prediction intervals represent the predicted intervals for the true value of the team's expected goals scored per match, based on each of the four separate advanced regression techniques that yielded the highest accuracies and adjusted R^2 values. After analyzing these prediction intervals–both for the true value of expected goals, and for the bootstrapped test dataset–it's confirmed that these machine learning methods are strong, and that the variables chosen are significantly important indicators for predicting expected goals.

DISCUSSION

This section details the results and conclusions based on the data and analyses provided above, including the most important variables in predicting expected goals, and what this analysis means for the broader soccer and sports analytics communities.

In analyzing this data, a plethora of different models, ranging from general multiple linear regression method to complex regression methods, were used in determining the variables that were most impactful in predicting expected goals. Among the most advanced modeling techniques–namely, Ridge, Lasso, Elastic Net, and Group Lasso regression–there was very little difference in the success of each model in predicting for expected goal values. Each of these four models produced very similar results, including low error values, and an adjusted R^2 of close to 0.70. While the differences in error and R^2 values among these models is minimal, it does appear that Elastic Net regression is marginally better than the rest, as it produced lower error values (lower MSE, RMSE, and MAPE), and a higher adjusted R^2 value. The choice of an Elastic Net model means that it was most appropriate to penalize non-significant variables to play a lesser impact in the prediction of the square root value of expected goals. This superior model ultimately showed that there were several key variables, which are discussed below. The results of Net Elastic regression, and other similar models, are significant enough to determine that these advanced regression techniques succeed in predicting the square-root transformed expected goal value for a given soccer team.

The modeling procedures discussed earlier are key in identifying the variables that are most important in predicting expected goals. In these advanced regression models, as well as earlier models, such as forward and backward stepwise regression, there are four variables that continually are deemed to be important. These four variables include: percentage of shots on target, number of crosses, number of touches in the penalty area, and number of recoveries. In particular, the number of touches in the penalty area and the number of recoveries are especially significant, as they had the lowest p-values among the four variables from forward and backward stepwise regression, and also were the only two predictor variables remaining (alongside the intercept) in advanced techniques such as Lasso, Elastic Net, and Group Lasso.

These four variables were found to be the most impactful in determining a team's predicted expected goal value. By maximizing the number of goals that a team is expected to score in a match, there is a greater chance that the team will end up winning. Therefore, there is great potential for team success in training with a focus on taking quality shots on goal more often, by crossing the ball into the box from the wings, putting the ball into the box more often, and playing a press on defense, to record higher numbers of recoveries. Practicing these concepts in drills, to maximize the

percentage of shots on target, number of crosses, number of touches in the penalty area, and the number of recoveries ultimately is important in helping overall team success in a match. All four variables have a positive relationship with expected goals, indicating that an increase in any will help the team (be expected to) score more often. When combined with strong defense, a potent offense that threatens many goals is very powerful. A tactical focus on getting the most from your players, including focusing on these specific parts of the match, have significant ramifications in setting a team up for the most success possible.

CONCLUSIONS

The effectiveness of these models represents another step taken to continue to analyze sports and its data in a deeper manner. The largely untapped potential of statistics in sports is boundless. Teams are finally beginning to realize that having a focus on analytics can be revolutionary to the success of the team, in matches, in maintaining physical health, in marketing, in making money, in increasing fan engagement, and much more. Soccer clubs are beginning to employ analytics teams to find the angles to get any advantage possible in each match. An extra advantage can be the difference between winning or losing, and ultimately between winning trophies or nothing at all.

The results of this research can impact how teams prepare for matches and strategize to win. By determining the variables that are most significant in predicting expected goals, there are clear metrics for teams to focus on maximizing. Achieving higher levels of expected goals has a direct correlation to actual goals scored in a match, and therefore determining how to get the best out a team can be related to these several key predictor variables. Additionally, it is clear how complex statistical models can be used to represent a wealth of sports data. As mentioned in the introduction, shots only represent a very minimal proportion of the total events occurring in a soccer match. However, using machine learning models, it is possible to account for a much greater proportion of the overall number of events occurring in a match, beyond considering only the total number of shots. This modeling enables a team to make decisions on a greater amount of data than only one or two unique predictor variables.

This study had several limitations, that would be key focus areas in the expansion of this work. Firstly, it would have been relevant to consider additional predictor variables–especially those that are more defensive in nature, or reflect transition periods in the match. It would also be beneficial to collect a larger dataset, expanding just beyond one team's records, and rather opening the door to expand into other teams, or even more broad, other international leagues. This could elicit insight on how different variables are of importance across borders, reflecting unique styles of play. With this larger dataset, a truly random train and test split could be used, further benefiting the study. The implementation of these changes would be exciting and important for future continuation of the study.

Advances in the world of sports analytics are very exciting, as they open the door to further innovation in a relatively new industry. New measures for team success are being found, and have the ability to continue to change the way players, coaches, and fans think about and analyze matches. This study reflects the growth of analytics opportunities in sports, and particularly soccer, and how much more can be discovered through deep learning methods. Key variables to improve team performance and health, and routes to commercial success are imperative for sports franchises around the world, and there is an opening to make meaningful change in this community. Our goal in writing this paper is to be a part of the sports analytics revolution. To push this research forward, we plan to extend the work to the spectrum of more sophisticated machine learning methods, such as deep learning (DL), neural networks (NN), and artificial neural networks (ANN) on big data related to professional soccer leagues.

ACKNOWLEDGMENTS

This study was made possible through the support of several parties, and we are sincerely grateful for their contributions. We extend our heartfelt thanks to the reviewers and the editorial team for their invaluable feedback, which significantly enhanced the quality of this manuscript.

DATA AND CODES AVAILABILITY

The data and R code are available upon request from the corresponding author.

CONFLICTS OF INTEREST

The authors have no known conflicts of interest in this research and manuscript submitted to the American Journal of Undergraduate Research (AJUR).

REFERENCES

1. Kuper, S., & Szymanski, S. (2018). *Soccernomics: Why England loses; Why Germany, Spain, and France Win; and Why One Day Japan, Iraq, and the United States Will Become Kings of the World's Most Popular Sport*. New York, NY, Nation Books.
2. Lewis, M. (2004). *Moneyball*. WW Norton.
3. Strategic Market Research. (2022, April). Industry Report and Statistics (Facts & Figures) - Number of Deployments & New Installations, Installation Cost & Demand by Sports Type. Sports Analytics Market Size, Global Trend Forecast – 2030. Retrieved January 31, 2023. *https://www.strategic marketresearch.com/market-report/sports-analytics-market*
4. Schilling, D. (2016, May 25). Sports analytics don't exist in "BlackWorld"? That's ridiculous. *https://www.theguardian.com/sport/blog/2016/may/25/sports-analytics-african-americans-michael-wilbon-article*
5. Chotiner, I. (2019, June 6). Jalen Rose has a Problem with Basketball Analytics. *https://www.newyorker.com/news/q-and-a/jalen-rose-has-a-problem-with-basketball-analytics*
6. Kirshner, A. (2022, December 15). Football has Found its New Bogeyman. *https://www.theatlantic.com/technology/archive/2022/11/sports-analytics-nfl-football/672275/*
7. Sommi, G. (2020, December 2). A Crash Course in Soccer Analytics. Retrieved February 16, 2023. *https://www.samford.edu/sports-analytics/fans/2020/A-Crash-Course-in-Soccer-Analytics*
8. Eggels, H. P. H. (2016). Expected Goals in Soccer: Explaining Match Results using Predictive Analytics (thesis). Eindhoven University of Technology, Eindhoven. *https://pure.tue.nl/ws/files/46945853/855660-1.pdf*
9. Statsbomb. (2022, November 15). *What is xG? How is it calculated? – Statsbomb – Data Champions*. What Are Expected Goals (xG)? Retrieved February 2, 2023. *https://statsbomb.com/soccer metrics/expected-goals-xg-explained/#:~:text=Put%20simply%2C%20Expected%20Goals%20(xG,scale%20between%200%20and%201*
10. Muller, J. (2022, March 4). Penalties are too generous a reward. we have a solution... and it involves running. *https://theathletic.com/3161748/2022/03/04/penalties-are-too-generous-a-reward-we-have-a-solution-and-it-involves-running/*
11. Tippett, J. (2019). *The Expected Goals Philosophy*. Self-published.
12. Perl, J., Grunz, A., & Memmert, D. (2013). Tactics Analysis in Soccer – An Advanced Approach. *International Journal of Computer Science in Sport*, 12(1), 33–44.
13. Muller, J. (2021, September 14). Possession Is The Puzzle Of Soccer Analytics. These Models Are Trying To Solve It. Retrieved February 1, 2023. *https://fivethirtyeight.com/features/possession -is-the-puzzle-of-soccer-analytics-these-models-are-trying-to-solve-it/*
14. Sánchez Gálvez, A. M., Álvarez González, R., Sánchez Gálvez, S., & Anzures García, M. (2022). Model to Predict the Result of a Soccer Match Based on the Number of Goals Scored by a Single Team. *ComputaciÃ¸n y Sistemas* 26(1). *https://doi.org/10.13053/cys-26-1-4172*
15. Decroos T., & Davis, J. (2019). Interpretable Prediction of Goals in Soccer. *http://statsbomb.com/wp -content/uploads/2019/10/decroos-interpretability-statsbomb.pdf*
16. Inan, T. (2021). Using Poisson model for goal prediction in European football. *Journal of Human Sport and Exercise*, 16(4), 942-955. *https://doi.org/10.14198/jhse.2021.164.16*
17. Liu, D. Z., Bratko, A., Prevc, J., Pataky, L., Noori, M., Sathyanarayana, S., & Lipovsek, U. (2020, December

8). Predicting Soccer Goals in Near Real Time Using Computer Vision. *https://aws.amazon.com/blogs/machine-learning/predicting-soccer-goals-in-near-real-time-using-computer-vision/*

18. Tweedale, A. (2022, June 24). Expected goals: Explained. *https://www.coachesvoice.com/cv/expected-goals-xg-explained/#: :text=Although%20the%20expected%20goals%20metric,such%20as%20an %20entire%20season*
19. Whitmore, J. (2023, August 8). What is expected goals (XG)? *https://theanalyst.com/na/2023/08/what-is-expected-goals-xg/*
20. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019, October 28). A public data set of spatio-temporal match events in soccer competitions. *https://www.nature.com/articles/s41597-019-0247-7*
21. Soccerment Research. (2019, November 21). AZ Alkmaar: Creating success through Data Analytics. *https://soccerment.com/az-alkmaar-success-data-analytics/*
22. Johnson, R. A., & Wichern, D. W. (2020). *Applied Multivariate Statistical Analysis* 6th ed. Pearson.
23. LaMorte, W. W. (2016, May 31). The Multiple Linear Regression Equation. *https://sphweb.bumc.bu.edu /otlt/mph-modules/bs/bs704-ep713_multivariablemethods/bs704-ep713_multivariablemethods2.html*
24. Mendenhall, W., & Sincich, T. (2011). In *A Second Course in Statistics: Regression Analysis* 7th ed., p. 333. Pearson.
25. Penn State Eberly College of Science. (n.d.). Detecting Multicollinearity Using Variance Inflation Factors. *https://online.stat.psu.edu/stat462/node/180/*
26. University of Cambridge Cognition and Brain Sciences Unit. (2003, November 27). *Collinearity: Its origins, effects, signs, symptoms and cures*. FAQ/Collinearity - CBU statistics Wiki. *https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Collinearity*
27. Dean, A., Voss, D., & Draguljić, D. (2017). *Design and Analysis of Experiments* 2nd ed. Springer.
28. Patrone, C. (2022, August 11). The Likelihood Ratio Test. *https://towardsdatascience.com/the-likelihood-ratio-test-463455b34de9*
29. Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications* 7th ed. Thomson Brooks/Cole.
30. Walczak, B., & Massart, D. L. (Eds.). (2000). *Data Handling in Science and Technology* Vol. 22. Elsevier.
31. Zajic, A. (2022, November 29). What is Akaike Information Criterion (AIC)? *https://builtin.com/data-science/what-is-aic*
32. Zach. (2021, August 25). A Complete Guide to Stepwise Regression in R - Statology. Retrieved February 16, 2023. *https://www.statology.org/stepwiseregression-r/*
33. Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.
34. Kenny, A., Solomon, D., Patwary, M. S., & Das, K. P. (2022). Sparse Regression with Clustered Predictors. *Journal of Statistical Research*, 56(1), 37–53. *https://doi.org/10.3329/jsr.v56i1.63945*
35. Friedman, J., Hastie, T., & Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, Articles 33 (1), 1–22. *https: //doi.org/10.18637/jss.v033.i01.*
36. Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55–67. *https://doi.org/10.1080/00401706.1970.10488634*
37. Penn State Eberly College of Science. (n.d.). Ridge Regression. *https://online.stat.psu.edu/stat857/node/155/*
38. Zach. (2020, November 13). Lasso Regression in R (Step-by-Step). Retrieved February 16, 2023. *https://www.statology.org/lasso-regression-in-r/*
39. Zou H. & Hastie T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society* B, 67(2), 301–320.
40. Yuan, M. and L. Lin (2006), Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society*, Series B 68, 49–67.
41. Lee, S.-H. (2023, January 12). Exclusive Lasso and group Lasso using R code. *https://ibkrcampus.com /ibkr-quant-news/exclusive-lasso-and-group-lasso-using-r-code/#: :text= While%20group%20lasso %20selects%20all,each%20other%20within%20each%20group*

42. Meier, L., van de Geer, S., & Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1), 53-âĂŞ71. *https://doi.org/10.1111/j.1467-9868.2007.00627.x*

43. Hodson, T. O. (2022, July 19). Root-mean-square error (RMSE) or mean absolute error (mae): When to use them or not. *https://gmd.copernicus.org/articles/15/5481/2022/*

44. Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician*, 69(4), 371–386. *https://doi.org/10.1080 /00031305.2015.1089789*

45. Penn State Eberly College of Science. (n.d.). Bootstrapping Methods. *https://online.stat.psu.edu/stat500/lesson/11/11.2/11.2.1*

ADDITIONAL REFERENCES

Arora, A. (2022, July 15). The Most Important Things You Need To Know About Elastic Net. *https://analyticsarora.com/the-most-important-things-you-need-to-know-about-elastic-net/*

Box, G. E. P., and Cox, D. R. (1964). An Analysis of Transformations (With Discussion). *Journal of the Royal Statistical Society B*, 26, 211–252.

Datasciencebeginners. (2020, May 17). *https://www.r-bloggers.com/2020/05/simple-guide-to-ridge-regression-in-r/#: :text=Overview,and%20tries%20to%20minimize%20them*

Glen, S. (2023, April 17). Akaike's Information Criterion: Definition, Formulas. *https://www.statisticshowto.com/akaikes-information-criterion/*

Hastie, T. J. & Pregibon, D. (1992). Generalized linear models. Chapter 6 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Heuer A. & Rubner O. (2012). How Does the Past of a Soccer Match Influence Its Future? Concepts and Statistical Analysis. *PLoS ONE* 7(11). 10.1371/journal.pone.0047678

Maher, M. (1982) Modelling association football scores. *Statist Neerland* 36: 109.

National Institute of Standards and Technology, U.S. Department of Commerce. (n.d.-a). Likelihood Ratio Tests. *https://www.itl.nist.gov/div898/handbook/apr/section2/apr233.htm*

Nopp S., Memmert D., Kempe M., & Vogelbein, M. (2012). Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer. *International Journal of Sports Science*. German Sport University: Cologne. 10.5923/s.sports.201401.05

Penn State Eberly College of Science. (n.d.). The Lasso. *https://online.stat.psu.edu/stat508/lesson/5/5.4*

Rue H., & Salvesen O. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician* 49, 399–418.

The Investopedia Team, Potters, C., & Eichler, R. (2023, September 28). R-Squared vs. Adjusted R-Squared: What's the Difference? *https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp#: :text=What%20Is%20the%20Difference%20Between,and%20R%2Dsquared%20does%20not*

Tianqi Chen and Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Van Roy M., Yang W., De Raedt L. & Davis J. (2021). Analyzing Learned Markov Decision Processes using Model Checking for Providing Tactical Advice in Professional Soccer. *https://eos.cs.kuleuven.be/sites/eos.cs.kuleuven.be/files/Defensive-Tactical-Advice-IJCAI-AISA2021.pdf*

Wallstreetmojo Team, & Vaidya, D. (n.d.). Elastic Net. *https://www.wallstreetmojo.com/elastic-net/*

Zhou, Y., R. Jin, and S. Hoi (2010), Exclusive Lasso for Multi-task Feature Selection. *In International Conference on Artificial Intelligence and Statistics*, 988–995.

ABOUT THE STUDENT AUTHOR

Tristan Rumsey is an alumnus of Butler University, having graduated in May 2024. Tristan majored in Statistics, and had minors in Business Administration and Data Science. While at school, and as a requirement of graduating with honors, Tristan wrote a thesis under the same name, and is now looking to publish the work with a prestigious under-

graduate journal. Currently, Tristan works in the industry in downtown Indianapolis. Tristan worked on this research under supervision of Dr. Shaha Patwary who is a co-author of this research.

PRESS SUMMARY

This work discusses the growth of analytics in sports within a soccer context. Specifically, the "expected goal" (xG), a key metric in modern soccer analytics, is the focus of the study, which seeks to analyze the most important predictor variables in maximizing xG. This analysis will give us key insights into how a team can practice, strategize, and plan appropriately to give themselves the best chance of winning, even if they might not have the same resources as larger clubs or organizations.

Print Edition ISSN 1536-4585
Online Edition ISSN 2375-8732